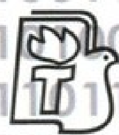
KHOA HỌC KHÁM PHÁ

Viktor Mayer-Schönberger
và Kenneth Cukier

Big Data
A revolution that will transform how we live, work, and think

DỮ LIỆU LỚN

Cuộc cách mạng sẽ làm thay đổi cách chúng ta sống, làm việc và tư duy



NHÀ XUẤT BẢN TRẺ

Viktor Mayer-Schönberger
và Kenneth Cukier

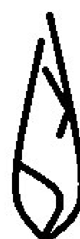
Vũ Duy Mẫn *dịch*

Big Data
A revolution that will transform how we live, work, and think

DỮ LIỆU LỚN

Cuộc cách mạng sẽ làm thay đổi cách chúng ta sống, làm việc và tư duy





NHÀ XUẤT BẢN TRẺ

Thông tin sách

Tên sách: **Dữ liệu lớn (Tủ sách Khoa học Khám phá)**

Nguyên tác: **Big data**

Tác giả: **Viktor Mayer-Schonberger, Kenneth Cukier**

Người dịch: **Vũ Duy Mẫn**

Nhà phát hành: **NXB Trẻ**

Nhà xuất bản: **NXB Trẻ**

Khối lượng: **350g**

Kích thước: **14.5 x 20.5 cm**

Ngày phát hành: **344**

Số trang: **03/2014**

Giá bìa: **120.000đ**

Thể loại: **Khoa học Khám phá**

Thông tin ebook

Thực hiện ebook: **thanhbt**

Ngày hoàn thành: **22/08/2017**

Dự án ebook #303 thuộc Tủ sách BOOKBT 

Giới thiệu

Màu sơn nào có thể cho bạn biết một chiếc xe đã qua sử dụng vẫn còn trong tình trạng tốt? Làm thế nào các công chức ở thành phố New York có thể xác định các hố ga nguy hiểm nhất trước khi chúng phát nổ? Và làm thế nào những cuộc tìm kiếm của Google dự đoán được sự lây lan của dịch cúm H1N1? Chìa khóa để trả lời những câu hỏi này, và nhiều câu hỏi khác, là dữ liệu lớn. "Dữ liệu lớn" đề cập đến khả năng đang phát triển của chúng ta để nắm giữ các bộ sưu tập lớn thông tin, phân tích, và rút ra những kết luận đôi khi sâu sắc đáng ngạc nhiên.

Lĩnh vực khoa học đang nổi lên này có thể chuyển vô số hiện tượng - từ giá vé máy bay đến các văn bản của hàng triệu cuốn sách - thành dạng có thể tìm kiếm được, và sử dụng sức mạnh tính toán ngày càng tăng của chúng ta để khám phá những điều chúng ta chưa bao giờ có thể nhìn thấy trước. Trong một cuộc cách mạng ngang tầm với Internet hoặc thậm chí in ấn, dữ liệu lớn sẽ thay đổi cách chúng ta nghĩ về kinh doanh, y tế, chính trị, giáo dục, và sự đổi mới trong những năm tới. Nó cũng đặt ra những mối đe dọa mới, từ sự kết thúc không thể tránh khỏi của sự riêng tư cho đến khả năng bị trừng phạt vì những thứ chúng ta thậm chí còn chưa làm, dựa trên khả năng của dữ liệu lớn có thể dự đoán được hành vi tương lai của chúng ta.Trong tác phẩm thông tuệ tuyệt vời và gây nhiều ngạc nhiên này, hai chuyên gia hàng đầu giải thích dữ liệu lớn là những gì, nó sẽ làm thay đổi cuộc sống của chúng ta như thế nào, và những gì chúng ta có thể làm để bảo vệ chính mình khỏi các mối nguy hiểm của nó. Dữ liệu lớn là cuốn sách lớn đầu tiên về điều to lớn sắp diễn ra. Bạn

đọc có thể quét các QR Code bên trong sách và trên bìa sách để xem các đoạn phim minh họa.

Tặng B và V V.M.S.Tặng cha mẹ của tôi

K.N.C.

1. HIỆN TẠI

NĂM 2009 MỘT VI-RÚT CÚM mới được phát hiện. Kết hợp các yếu tố của các vi-rút gây cúm gà, chủng mới này, được gọi là H1N1, đã lây lan nhanh chóng. Trong vài tuần, các cơ sở y tế khắp thế giới lo sợ một đại dịch khủng khiếp đang xảy ra. Một số nhà bình luận đã cảnh báo về một dịch bệnh có quy mô của dịch cúm Tây Ban Nha vào năm 1918, lây nhiễm cho nửa tỷ người và làm chết hàng chục triệu người. Tồi tệ hơn là không hề có vắc-xin nào để chống lại vi-rút mới này. Hy vọng duy nhất của cơ quan y tế là giảm mức lây lan. Nhưng để làm điều đó, họ cần biết bệnh đã lan tới đâu.

Ở Mỹ, Trung tâm Kiểm soát và Phòng chống Bệnh dịch (CDC) đã yêu cầu các bác sĩ thông báo về các ca bệnh cúm mới. Nhưng bức tranh thật về đại dịch vẫn luôn bị chậm trễ một hoặc hai tuần. Nhiều người có thể bị bệnh vài ngày rồi mới đi gặp bác sĩ. Việc chuyển tiếp thông tin về các cơ quan trung ương đòi hỏi thời gian, và CDC chỉ xử lý các con số mỗi tuần một lần. Với một bệnh dịch lây lan nhanh, hai tuần chậm trễ cũng giống như dài vô tận. Sự chậm trễ này đã hoàn toàn vô hiệu hóa các cơ quan y tế tại những thời điểm gay cấn nhất.

Lúc việc đó xảy ra, vài tuần trước khi vi-rút H1N1 xuất hiện rầm rộ trên các phương tiện truyền thông, các kỹ sư của công ty Internet khổng lồ Google đã đăng một bài đáng chú ý trên tạp chí khoa học *Nature*. Nó đã tạo một chuyện giật gân trong giới chức y tế và các nhà khoa học máy tính, nhưng ngoài ra thì ít được quan tâm. Các tác giả lý giải Google có thể "dự đoán" sự lây lan của bệnh cúm mùa đông ở Mỹ như thế nào, không chỉ ở mức độ toàn quốc, mà còn chi tiết tới mức vùng và thậm chí tới mức tiểu bang. Google có thể đạt được điều này bằng cách xem xét những gì người sử dụng đã tìm kiếm trên Internet. Bởi Google nhận được hơn ba tỷ câu hỏi tìm kiếm mỗi ngày và lưu giữ tất cả chúng, nên nó có vô số dữ liệu để phân tích.

Google lấy 50 triệu cụm từ được tìm kiếm phổ biến nhất của người Mỹ và so sánh chúng với dữ liệu của CDC về sự lây lan của bệnh cúm mùa giữa các năm 2003 và 2008. Ý tưởng là để xác định các khu vực bị lây nhiễm vi-rút cúm thông qua những gì người ta tìm kiếm trên Internet, và không ai khác có nhiều dữ liệu, năng lực tính toán và hiểu biết về thống kê như Google.

Dù các chuyên viên của Google phỏng đoán các lệnh tìm kiếm có thể nhằm thu lượm thông tin về cúm - gõ các câu đại loại như "thuốc ho và sốt" - nhưng không phải vậy: họ không biết, và họ đã thiết kế một hệ thống không quan tâm tới điều đó. Tất cả những gì hệ thống của họ làm là phát hiện mối tương quan giữa tần suất của một số câu hỏi tìm kiếm và sự lây lan của bệnh cúm theo thời gian và không gian. Tổng cộng, họ xử lý một lượng đáng kinh ngạc 450 triệu mô hình toán học khác nhau để kiểm tra các điều kiện tìm kiếm, so sánh các dự đoán của họ với các trường hợp bệnh thực tế từ CDC trong năm 2007 và 2008. Và họ đã vớ được vàng: phần mềm của họ tìm thấy một sự kết hợp của 45 điều kiện tìm kiếm mà khi sử dụng cùng với một mô hình toán học, có một mối tương quan mạnh mẽ giữa phỏng đoán của họ và các số liệu chính thức trên toàn quốc. Giống như CDC, họ có thể cho biết cúm đã lây lan tới đâu, nhưng khác với CDC, họ có thể nói điều đó gần như trong thời gian thực, chứ không phải trễ một hoặc hai tuần.

Do vậy, khi dịch bệnh H1N1 tấn công vào năm 2009, hệ thống của Google đã chứng tỏ là một chỉ báo có ích hơn và nhanh hơn so với các số liệu thống kê của chính phủ thường chậm trễ. Các quan chức y tế đã được trang bị những thông tin có giá trị.

Điều gây ấn tượng là phương pháp của Google không liên quan gì đến việc phân phối gạc miệng hoặc liên hệ với các phòng khám. Thay vào đó, nó được xây dựng trên "dữ liệu lớn" - khả năng của xã hội khai thác thông tin theo những cách thức mới để đưa ra

những kiến thức hữu ích hay những sản phẩm và dịch vụ có giá trị đáng kể. Với nó, khi đại dịch kế tiếp xảy ra, thế giới sẽ có sẵn một công cụ tốt hơn để dự đoán và do đó ngăn chặn sự lây lan.



Phim minh họa phương pháp của Google

Y tế công chỉ là một lĩnh vực trong đó dữ liệu lớn đang làm nên một sự khác biệt vĩ đại. Nhiều lĩnh vực khác cũng đang được định hình lại bởi dữ liệu lớn. Dịch vụ mua vé máy bay là một thí dụ.

Năm 2003, Oren Etzioni cần bay từ Seattle tới Los Angeles để dự lễ cưới em trai của ông. Nhiều tháng trước đó, ông lên mạng và mua một vé máy bay, tin rằng càng mua sớm, vé càng rẻ. Trên chuyến bay, do tò mò, Etzioni hỏi người ngồi kế bên xem giá vé của ông ta là bao nhiêu và ông ta mua khi nào. Hóa ra ông ta trả thấp hơn nhiều so với Etzioni, mà thậm chí ông ta mới chỉ mua vé gần đây. Khá tức giận, Etzioni hỏi một hành khách khác và

một hành khách khác nữa. Hầu hết họ đã trả ít tiền hơn.

Với hầu hết chúng ta, ý nghĩa của cảm giác bị lừa có thể đã tiêu tan khi chúng ta gấp khay bàn ăn trước mặt, dựng thẳng ghế và khóa thắt lưng an toàn. Nhưng Etzioni là một trong những nhà khoa học máy tính hàng đầu của Mỹ. Ông nhìn thế giới như một chuỗi các bài toán dữ-liệu-lớn có thể giải được. Và ông đang làm chủ chúng từ khi là người đầu tiên tốt nghiệp Đại học Harvard về chuyên ngành khoa học máy tính vào năm 1986.

Từ căn phòng của mình tại Đại học Washington, ông đã khởi xướng những công ty dữ-liệu-lớn trước khi thuật ngữ "dữ liệu lớn" được biết tới. Ông đã giúp phát triển một trong những công cụ tìm kiếm Web đầu tiên, MetaCrawler, được đưa ra sử dụng vào năm 1994 rồi sau được bán cho InfoSpace, lúc đó là một công ty bất động sản trực tuyến lớn. Ông đã đồng sáng lập Netbot, trang web mua hàng so sánh lớn đầu tiên, sau đó bán nó cho Excite. Ông khởi động công ty làm công cụ trích ý nghĩa từ các văn bản, gọi là ClearForest, sau này được Reuters mua lại.

Trở lại câu chuyện chính, Etzioni quyết tìm ra cách để có thể biết liệu một giá vé ta thấy trên mạng có phải là một giá tốt hay không. Một chỗ ngồi trên máy bay là một thương phẩm: mỗi chỗ về cơ bản là hoàn toàn giống với những chỗ khác trên cùng chuyến bay. Nhưng giá lại rất khác nhau, dựa trên vô số yếu tố mà chủ yếu chỉ chính các hãng bay mới biết.

Etzioni đi đến kết luận ông không cần giải mã ý nghĩa hay nguyên nhân giá cả khác nhau. Thay vào đó, ông đơn giản phải dự đoán liệu giá được báo có khả năng tăng hay giảm trong tương lai. Điều này là khả thi, nếu không nói là dễ. Những gì cần thiết là phân tích tất cả các vé bán cho một tuyến đường và khảo sát các giá phải trả tương quan với số ngày mua trước lúc khởi hành.

Nếu giá trung bình của vé có xu hướng giảm, thì rất nên đợi để mua sau. Nếu giá trung bình có xu hướng tăng, hệ thống sẽ khuyến cáo mua vé ngay với giá được báo. Nói cách khác, thứ cần thiết là một dạng cải tiến của cuộc điều tra thông tin Etzioni đã thực hiện trên tầng cao 30.000 feet. Chắc chắn đó là một bài toán lớn khác của khoa học máy tính. Nhưng, đó là một bài toán ông có thể giải được. Do vậy Etzioni đã bắt tay vào công việc.

Sử dụng một mẫu gồm 12.000 lượt thống kê giá vé qua “thu lượm” thông tin trong 41 ngày từ một trang web du lịch, Etzioni đã tạo được một mô hình dự báo giúp hành khách tiết kiệm chi phí. Mô hình không có hiểu biết về câu hỏi *tại sao*, mà chỉ biết về câu hỏi *cái gì.* Nó không biết bất kỳ tham biến nào tham gia vào những quyết định về giá của các hãng hàng không, chẳng hạn số chỗ ngồi còn chưa bán được, mùa vụ, hay một loại thu xếp lưu trú qua đêm thứ Bảy có thể làm giảm giá vé. Hệ thống dự đoán dựa vào những gì đã biết: xác suất có được từ những chuyến bay khác. “Mua hay không mua, đó là câu hỏi”, Etzioni ngẫm nghĩ. Thế nên ông đặt tên rất thích hợp cho dự án là Hamlet.

Dự án nhỏ đã phát triển thành một doanh nghiệp khởi động được hỗ trợ bằng vốn mạo hiểm mang tên Farecast. Bằng cách dự báo giá của một vé máy bay rất có thể tăng hoặc giảm, và tăng hoặc giảm bao nhiêu, Farecast trao quyền cho người tiêu dùng lựa chọn khi nào thì nhấp vào nút “mua”. Nó trang bị cho họ thông tin mà trước đây họ chưa bao giờ truy cập được. Đề cao tính tự minh bạch, Farecast cho điểm độ tin cậy đối với dự báo của chính nó và cũng thông báo số điểm này cho người sử dụng.

Để hoạt động, hệ thống cần rất nhiều dữ liệu. Nhằm cải thiện hiệu suất của hệ thống, Etzioni đã nhúng tay vào một trong các cơ sở dữ liệu đăng ký chỗ của ngành hàng không. Với thông tin này, hệ thống có thể đưa ra các dự báo dựa vào từng chỗ ngồi trên từng chuyến bay cho hầu hết các tuyến bay của hàng không thương mại Mỹ trong một năm. Farecast xử lý gần 200 tỷ bản ghi

giá vé máy bay để đưa ra các dự báo của nó. Làm như vậy, Farecast đã tiết kiệm được cho người tiêu dùng bộn tiền.

Với mái tóc màu nâu cát, nụ cười chân thành, và nét đẹp hiền hậu, Etzinoni hầu như không có vẻ là loại người có thể phủ nhận hàng triệu đôla doanh thu tiềm năng của ngành hàng không. Trong thực tế, ông đặt tầm ngắm của mình còn xa hơn thế. Năm 2008 ông đặt kế hoạch áp dụng phương pháp này cho các sản phẩm khác như phòng khách sạn, vé nghe hòa nhạc, và xe hơi cũ: tất cả mọi thứ với sự khác biệt rất ít về sản phẩm, có độ biến động giá cả cao, và có rất nhiều dữ liệu. Nhưng trước khi ông có thể triển khai được các kế hoạch của mình, Microsoft đã tới gõ cửa, mua Farecast với khoảng $110 triệu, và tích hợp nó vào công cụ tìm kiếm Bing. Tới năm 2012 hệ thống đã khuyến cáo đúng tới 75% và tiết kiệm cho hành khách trung bình $50 mỗi vé.

Farecast là hình ảnh thu nhỏ của một công ty dữ-liệu-lớn và một thí dụ cho thấy thế giới hướng tới đâu. Etzioni không thể thiết lập công ty năm hoặc mười năm sớm hơn. “Đó là điều bất khả”, ông nói. Lượng sức mạnh tính toán và lưu trữ cần thiết đã là quá lớn. Nhưng, mặc dù những thay đổi về công nghệ là yếu tố quan trọng giúp cho nó trở thành khả thi, một số điều quan trọng hơn cũng thay đổi - những điều tinh tế. Đã có sự thay đổi trong suy nghĩ về việc dữ liệu có thể được sử dụng như thế nào.

Dữ liệu không còn được xem là tĩnh hoặc cũ, tính hữu ích của dữ liệu kết thúc một khi mục tiêu mà vì nó dữ liệu được thu thập đã đạt được, chẳng hạn sau khi máy bay đã hạ cánh (hoặc trong trường hợp của Google, khi một câu hỏi tìm kiếm đã được xử lý). Thay vào đó, dữ liệu trở thành một nguyên liệu thô của doanh nghiệp, một đầu vào kinh tế quan trọng, được sử dụng để tạo ra một dạng mới của giá trị kinh tế. Thực tế, với suy nghĩ đúng đắn, dữ liệu có thể được dùng lại một cách thông minh để trở thành một suối nguồn của thông tin và những dịch vụ mới. Dữ liệu có

thể tiết lộ bí mật cho những ai có sự khiêm nhường, sự sẵn lòng và công cụ để lắng nghe.

Hãy để cho dữ liệu nói

Thật dễ nhận thấy những thành quả của xã hội thông tin, với một điện thoại di động và một máy tính bỏ túi mỗi người, cùng các hệ thống công nghệ thông tin lớn trong văn phòng khắp mọi nơi. Nhưng điều người ta ít thấy rõ hơn là chính thông tin.

Một nửa thế kỷ sau khi máy tính bước vào xã hội chính thống, dữ liệu bắt đầu được tích lũy nhiều tới mức mà một điều gì đó mới mẻ và đặc biệt sắp xảy ra. Không những thế giới tràn ngập thông tin nhiều hơn bao giờ hết, mà thông tin còn tăng nhanh hơn. Sự thay đổi về quy mô đã dẫn đến một sự thay đổi về trạng thái. Thay đổi về lượng đã dẫn tới thay đổi về chất. Các khoa học như thiên văn, gen, mới được trải nghiệm sự bùng nổ trong những năm 2000, đã đưa ra thuật ngữ “dữ liệu lớn”, khái niệm mà nay đã di trú vào tất cả các lĩnh vực của đời sống con người.

Không có một định nghĩa chính xác cho dữ liệu lớn. Ban đầu ý tưởng là dung lượng thông tin đã tăng quá lớn tới mức số lượng cần khảo sát không còn vừa vào bộ nhớ các máy tính dùng để xử lý, do vậy các kỹ sư cần cải tạo các công cụ họ dùng để có thể phân tích được tất cả thông tin. Đó là xuất xứ của các công nghệ xử lý mới như MapReduce của Google và nguồn mở tương đương của nó, Hadoop, khởi đầu từ Yahoo. Những công nghệ này cho phép ta quản lý những khối lượng dữ liệu lớn hơn nhiều so với trước đây, và quan trọng là không cần đưa dữ liệu vào các hàng ngăn nắp hoặc các bảng cơ sở dữ liệu cổ điển. Các công nghệ nghiền dữ liệu khác, bỏ qua các cấu trúc phân cấp và đồng nhất cứng nhắc cổ điển, cũng ở trong tầm ngắm. Đồng thời, do các công ty Internet có thể thu thập được vô số dữ liệu quý giá và có động cơ kinh tế lớn để khai thác chúng, nên các công ty này trở thành người sử dụng hàng đầu của các công nghệ xử lý hiện đại

nhất, vượt qua các công ty truyền thống, đôi khi có tới hàng chục năm kinh nghiệm nhiều hơn.

Một cách để suy nghĩ về vấn đề ngày hôm nay - và cũng là cách chúng tôi thực hiện trong cuốn sách này - là: dữ liệu lớn đề cập tới những thứ người ta có thể làm với một quy mô lớn mà không thể làm với một quy mô nhỏ hơn, để trích xuất những hiểu biết mới hoặc tạo ra những dạng giá trị mới, theo những cách thức có thể làm thay đổi các thị trường, các tổ chức, mối quan hệ giữa các công dân và các chính phủ, và hơn thế nữa.

Nhưng đó chỉ là bước khởi đầu. Thời đại của dữ liệu lớn thách thức cách chúng ta sống và tương tác với thế giới. Nổi bật nhất, xã hội sẽ phải cắt giảm một số nỗi ám ảnh của nó về quan hệ nhân quả để đổi lấy mối tương quan đơn giản, không biết *tại sao* mà chỉ biết *cái gì*. Điều đó làm đổ vỡ hàng thế kỷ các tập quán đã được thiết lập và thách thức hiểu biết cơ bản nhất của chúng ta về việc làm thế nào để đưa ra được quyết định và hiểu được thực tế.

Dữ liệu lớn đánh dấu bước khởi đầu của một biến đổi lớn. Giống như nhiều công nghệ mới, dữ liệu lớn chắc chắn sẽ trở thành nạn nhân của chu kỳ thổi phồng khét tiếng của Thung Lũng Silicon: sau khi được tiếp đón trên trang đầu của các tạp chí và tại các hội nghị công nghiệp, xu hướng này sẽ bị ruồng bỏ và rất nhiều công ty khởi động say mê dữ liệu sẽ bị lúng túng. Nhưng cả thái độ say mê và nguyền rủa đều hiểu lầm một cách khá sâu tầm quan trọng của những gì đang xảy ra. Đúng như kính thiên văn tạo điều kiện cho chúng ta hiểu biết được vũ trụ và kính hiển vi cho phép chúng ta hiểu biết được vi trùng, các kỹ thuật mới để thu thập và phân tích những tập hợp lớn dữ liệu sẽ giúp chúng ta tìm ra ý nghĩa của thế giới theo những cách thức mà chúng ta mới chỉ vừa bắt đầu ưa thích. Trong cuốn sách này, chúng tôi không hẳn là những kẻ truyền giáo của dữ liệu lớn mà chỉ là những người đưa tin. Và, một lần nữa xin nhấn mạnh, cuộc cách mạng

thật sự không phải ở những chiếc máy tính toán dữ liệu mà ở chính dữ liệu và cách ta sử dụng chúng.

Để đánh giá mức độ một cuộc cách mạng thông tin đã tiến triển tới đâu, ta hãy xem xét các xu hướng xuyên suốt các lĩnh vực của xã hội. Lấy ví dụ thiên văn học. Khi Sloan Digital Sky Survey (SDSS - Trạm quan sát bầu trời bằng kỹ thuật số Sloan) bắt đầu hoạt động vào năm 2000, kính thiên văn của nó tại New Mexico trong mấy tuần đầu tiên đã thu thập nhiều dữ liệu hơn những gì được thu thập trong toàn bộ lịch sử của ngành thiên văn. Đến năm 2010, lưu trữ của trạm đã bạt ngàn với con số khổng lồ 140 tera (10 mũ 12) byte thông tin. Nhưng kẻ kế nhiệm, kính thiên văn của Large Synoptic Survey (LSST) ở Chile, dự kiến vận hành vào năm 2016, cứ mỗi năm ngày sẽ thu thập được lượng dữ liệu tương đương như thế.

Những số lượng vô cùng to lớn như vậy cũng có thể được tìm thấy ngay xung quanh chúng ta. Khi các nhà khoa học lần đầu giải mã gen người vào năm 2003, họ đã mất một thập kỷ làm việc miệt mài để xác định trình tự cho ba tỷ cặp cơ sở. Bây giờ, sau một thập kỷ, một thiết bị đơn lẻ cũng có thể xác định trình tự cho số lượng DNA như vậy chỉ trong một ngày. Trong ngành tài chính, khoảng 7 tỷ cổ phiếu được mua bán mỗi ngày trên các thị trường chứng khoán Mỹ, trong số đó khoảng hai phần ba được giao dịch bằng các thuật toán máy tính dựa trên các mô hình toán học xử lý hàng núi dữ liệu để dự đoán lợi nhuận trong khi cố gắng giảm thiểu rủi ro.

Các công ty Internet đặc biệt bị tràn ngập. Google xử lý hơn 24 peta (10 mũ 15) byte dữ liệu mỗi ngày, một khối lượng gấp hàng ngàn lần tất cả các ấn phẩm trong Thư viện Quốc hội Mỹ. Facebook, một công ty không hề tồn tại một thập kỷ trước, nhận hơn 10 triệu ảnh mới được tải lên mỗi giờ. Các thành viên Facebook nhấp nút “like” hoặc gửi lời bình luận gần ba tỷ lần mỗi ngày, tạo một dấu vết số để công ty có thể “đào xới” nhằm biết

được các sở thích của người sử dụng. Trong khi đó, 800 triệu người sử dụng dịch vụ Youtube của Google tải lên hơn một giờ video mỗi giây. Thành viên của mạng Twitter tăng khoảng 200 phần trăm mỗi năm và đến năm 2012 đã có hơn 400 triệu *tweet* mỗi ngày.

Từ khoa học tới y tế, từ ngân hàng tới Internet, các lĩnh vực có thể khác nhau, nhưng cùng nhau chúng đều có một câu chuyện tương tự: số lượng dữ liệu trong thế giới đang tăng rất nhanh, vượt sức không chỉ những chiếc máy tính mà cả trí tưởng tượng của chúng ta.

Nhiều người đã thử đưa ra một con số thực tế về lượng thông tin xung quanh chúng ta và tính toán xem nó tăng như thế nào. Họ đã có những mức độ thành công khác nhau bởi họ đo lường những thứ khác nhau.. Một trong những nghiên cứu toàn diện hơn được Martin Hilbert của Trường Truyền thông và Báo chí Annenberg thuộc Đại học Nam California thực hiện. Ông đã nỗ lực đưa ra một con số cho mọi thứ đã từng được sản xuất, lưu trữ và truyền tải. Chúng không chỉ bao gồm sách, tranh, email, ảnh, nhạc, và phim (cả dạng analog và digital), mà còn có trò chơi điện tử, cuộc gọi điện thoại, thậm chí các hệ thống điều hướng xe và thư gửi qua bưu điện. Ông cũng bao gồm các phương tiện truyền thông phát sóng như truyền hình và radio, dựa trên tiếp cận khán giả.

Theo ước lượng của Hilbert, hơn 300 exa (10 mũ 18) byte dữ liệu lưu trữ đã tồn tại vào năm 2007. Để dễ hình dung ý nghĩa của nó, thử nghĩ thế này. Một bộ phim dài ở dạng kỹ thuật số có thể được nén vào một tập tin 1 giga byte. Một exa byte là 1 tỷ gi ga byte. Tóm lại là vô cùng nhiều. Điều thú vị là năm 2007 chỉ khoảng 7 phần trăm dữ liệu ở dạng analog (giấy, sách, ảnh in, vân vân). Phần còn lại là ở dạng digital - kỹ thuật số. Nhưng mới gần đây, bức tranh đã rất khác. Mặc dù những ý tưởng của cuộc “cách mạng thông tin” và “thời đại kỹ thuật số” đã xuất hiện từ những

năm 1960, chúng mới chỉ trở thành hiện thực ở vài khía cạnh. Tới tận năm 2000, mới chỉ có một phần tư thông tin lưu trữ của thế giới được số hóa. Ba phần tư còn lại vẫn ở trên giấy, phim, đĩa nhựa, băng từ, và những thứ tương tự.

Lượng thông tin kỹ thuật số lúc đó chưa nhiều - một điều thật kém cỏi với những ai lướt mạng và mua sách trực tuyến suốt thời gian dài. (Thực tế, vào năm 1986, khoảng 40 phần trăm sức mạnh tính toán thông dụng của thế giới là ở dạng những chiếc máy tính bỏ túi, lớn hơn sức mạnh của tất cả các máy tính cá nhân tại thời điểm đó.) Nhưng vì dữ liệu kỹ thuật số phát triển rất nhanh - cứ hơn ba năm lại tăng gấp đôi, theo Hilbert - nên tình hình đã nhanh chóng tự đảo ngược. Thông tin analog, ngược lại, không hề tăng. Do vậy vào năm 2013 lượng thông tin lưu trữ trong thế giới ước lượng khoảng 1.200 exa byte, trong đó chưa đến 2 phần trăm là phi kỹ thuật số.

Chẳng có cách nào phù hợp để hình dung kích thước như vậy của dữ liệu là có ý nghĩa gì. Nếu tất cả được in thành sách, chúng có thể phủ kín bề mặt của nước Mỹ với chiều dày 52 lớp. Nếu được ghi vào CD-ROM và xếp chồng lên nhau, chúng có thể tạo thành 5 cột vươn cao tới mặt trăng. Vào thế kỷ thứ ba trước Công nguyên, khi Ptolemy II của Ai Cập cố gắng lưu trữ một bản của mỗi tác phẩm từng được viết ra, Thư viện lớn của Alexandria đã tượng trưng cho toàn bộ tri thức của thế giới. Trận lũ lớn kỹ thuật số hiện đang quét qua trái đất tương đương với việc cung cấp cho mỗi người sống trên trái đất hôm nay 320 lần nhiều hơn thông tin như ước lượng đã được lưu trữ ở Thư viện Alexandria.

Mọi thứ thật sự được tăng tốc. Lượng thông tin lưu trữ tăng nhanh hơn gấp bốn lần so với kinh tế thế giới, trong khi sức mạnh xử lý của máy tính tăng nhanh hơn gấp chín lần. Không ngạc nhiên khi người ta than phiền tình trạng quá tải thông tin. Ai cũng bị tác động bởi những thay đổi này.

Hãy nhìn một cách dài hạn, bằng cách so sánh trận lũ dữ liệu hiện tại với cuộc cách mạng thông tin trước đây, cách mạng in sách của Gutenberg được phát minh khoảng năm 1439. Trong năm mươi năm từ 1453 tới 1503 khoảng 8 triệu cuốn sách đã được in, theo nhà sử học Elizabeth Eisenstein. số lượng đó được xem là nhiều hơn tất cả những gì các thầy thông giáo đã chép ra kể từ lúc thiết lập nên Constantinople khoảng 1.200 năm trước. Nói cách khác, phải mất năm mươi năm để khối lượng thông tin tăng gấp đôi ở Âu châu, so với khoảng mỗi ba năm hiện nay.

Sự tăng trưởng này có ý nghĩa gì? Peter Norvig, một chuyên gia trí tuệ nhân tạo tại Google, thích nghĩ về nó với một sự tương tự về hình ảnh. Ông yêu cầu chúng tôi trước hết xem xét con ngựa mang tính biểu tượng từ các hình vẽ hang động ở Lascaux, Pháp, từ thời Paleolithic khoảng 17 ngàn năm trước. Sau đó nghĩ tới một bức ảnh của một con ngựa - hoặc tốt hơn là những phác họa của Pablo Picasso, trông không khác mấy các hình vẽ hang động. Thực tế, khi Picasso được cho xem các hình vẽ Lascaux, từ thời đó ông đã hài hước rằng: “Chúng ta đã không phát minh ra được thứ gì cả”.

Những lời của Picasso đúng ở một mức độ, nhưng không đúng ở một mức độ khác. Hãy nhớ lại bức ảnh chụp con ngựa. Trong khi phải mất nhiều thời gian để vẽ hình một con ngựa, bây giờ sự miêu tả một con ngựa có thể được thực hiện nhanh hơn nhiều với nhiếp ảnh. Đó là sự thay đổi, nhưng có thể đó không phải là thứ quan trọng nhất, bởi về cơ bản nó vẫn thế: hình ảnh của một con ngựa. Nhưng bây giờ, Norvig yêu cầu thu hình của một con ngựa và tăng tốc độ lên tới 24 khung hình mỗi giây. Sự thay đổi lượng đã tạo ra một thay đổi chất. Một bộ phim khác một cách cơ bản với một bức ảnh tĩnh. Với dữ liệu lớn cũng vậy: bằng cách thay đổi số lượng, chúng ta thay đổi bản chất.

Hãy xét một sự tương tự từ công nghệ nano - nơi mà mọi thứ trở nên nhỏ đi, chứ không lớn lên. Nguyên lý phía sau của công nghệ

nano là khi đạt tới mức phân tử, các tính chất vật lý có thể thay đổi. Hiểu biết những đặc tính mới này có nghĩa là ta có thể sắp đặt để vật chất làm những thứ mà trước đây không thể làm được. Thí dụ, tại quy mô nano, kim loại có thể mềm dẻo hơn và gốm sứ có thể căng ra được. Ngược lại, khi tăng quy mô của dữ liệu, chúng ta có thể thực hiên được những thứ mới mà không thể nào thực hiện được khi chỉ làm việc với những số lượng nhỏ hơn.

Đôi khi những khó khăn mà chúng ta đang sống cùng thực ra chỉ là những chức năng của quy mô mà chúng ta hoạt động trong đó, và điều này cũng tương tự cho tất cả mọi thứ. Hãy xét một sự tương tự thứ ba, lại cũng từ các ngành khoa học. Đối với con người, định luật vật lý quan trọng nhất là lực hấp dẫn: nó ngự trị trên tất cả những gì chúng ta làm. Nhưng đối với những côn trùng nhỏ xíu, trọng lực hầu như vô nghĩa. Đối với một số loài như nhện nước, định luật vật lý có tác động với chúng chính là sức căng bề mặt, cho phép chúng đi qua một cái ao mà không chìm xuống.

Với thông tin, cũng như với vật lý, kích thước là quan trọng. Do đó, Google có thể xác định được sự lây lan của bệnh cúm chính xác như các dữ liệu chính thức dựa trên việc bệnh nhân thực sự tới gặp bác sĩ. Google có thể làm điều này bằng cách kết hợp hàng trăm tỷ từ khóa tìm kiếm - và nó có thể đưa ra một câu trả lời gần như trong thời gian thực, nhanh hơn nhiều các nguồn chính thức. Tương tự như vậy, Farecast của Etzioni có thể dự đoán sự biến động giá của một chiếc vé máy bay và do đó chuyển quyền lực kinh tế đáng kể vào tay người tiêu dùng. Nhưng cả hai chỉ có thể làm tốt như vậy bằng cách phân tích hàng trăm tỷ điểm dữ liệu.

Hai ví dụ trên cho thấy tầm quan trọng về khoa học và xã hội của dữ liệu lớn cũng như mức độ mà dữ liệu lớn có thể trở thành một nguồn giá trị kinh tế. Chúng đánh dấu hai cách thức mà thế giới dữ liệu lớn đã sẵn sàng để cải tổ tất cả mọi thứ, từ các doanh

nghiệp và các ngành khoa học tới chăm sóc sức khỏe, chính phủ, giáo dục, kinh tế, nhân văn, và mọi khía cạnh khác của xã hội.

Mặc dù mới chỉ đang ở buổi bình minh của dữ liệu lớn, chúng ta dựa vào nó hàng ngày. Bộ lọc thu rác được thiết kế để tự động thích ứng khi các loại email rác thay đổi: phần mềm không thể được lập trình để biết ngăn chặn “via6ra” hoặc vô số các biến thể của nó. Các trang web hẹn hò ghép các cặp trên cơ sở các thuộc tính tương quan thế nào với những cặp đã được ghép thành công trước đó. Tính năng “tự động sửa lỗi” trong điện thoại thông minh theo dấu các thao tác của chúng ta và bổ sung thêm những từ mới vào từ điển chính tả của nó dựa trên những gì chúng ta nhập vào. Tuy nhiên, những ứng dụng này mới chỉ là sự bắt đầu. Từ những chiếc xe hơi có thể phát hiện khi nào cần đi hướng khác hoặc phanh, đến máy tính Watson của IBM đánh bại con người trong trò chơi truyền hình *Jeopardy!*, cách tiếp cận này sẽ cải tạo nhiều khía cạnh của thế giới chúng ta đang sống.

Tại cốt lõi của nó, dữ liệu lớn là về các dự đoán. Mặc dù nó được mô tả như một phần của ngành khoa học máy tính được gọi là trí tuệ nhân tạo, và cụ thể hơn, một lĩnh vực được gọi là học qua máy, việc đặc trưng hóa này gây nhầm lẫn. Dữ liệu lớn không phải là về việc cố gắng “dạy” một máy tính “nghĩ” giống như con người. Thay vào đó, nó là về việc áp dụng toán học với số lượng lớn dữ liệu để suy ra xác suất: khả năng một email là thư rác; các ký tự gõ “teh” đáng lẽ phải là “the”; quỹ đạo và vận tốc của một người lái xe cho phép anh ta qua được phố đúng thời gian. Điều quan trọng là các hệ thống này thao tác tốt bởi chúng được nạp với rất nhiều dữ liệu để làm cơ sở cho các dự đoán của chúng. Hơn nữa, hệ thống được thiết kế để tự cải thiện theo thời gian, bằng cách giữ một nhãn (tab) về các tín hiệu và mẫu tốt nhất để tìm kiếm khi có thêm dữ liệu được đưa vào.

Trong tương lai - và có thể sớm hơn chúng ta nghĩ - nhiều khía cạnh của cuộc sống sẽ được tăng cường hoặc thay thế bằng

những hệ thống máy tính, những khía cạnh mà hôm nay là phạm vi hoạt động duy nhất của sự phán xét con người. Không chỉ việc lái xe hoặc mai mối, mà cả những việc phức tạp hơn. Rốt cuộc, Amazon có thể giới thiệu được cuốn sách lý tưởng, Google có thể xếp hạng được trang web phù hợp nhất, Facebook biết được sở thích của chúng ta, và Linkedln tiên đoán được người mà chúng ta biết. Cũng những công nghệ này sẽ được áp dụng cho chẩn đoán bệnh, đề xuất phương pháp điều trị, thậm chí có thể xác định "tội phạm" trước khi hắn thực sự phạm tội. Cũng giống như Internet hoàn toàn thay đổi thế giới bằng cách thêm truyền thông vào máy tính, dữ liệu lớn sẽ thay đổi các khía cạnh cơ bản của cuộc sống bằng cách cho nó một kích thước định lượng chưa hề có trước đây.

Nhiều hơn, lộn xộn, đủ tốt

Dữ liệu lớn sẽ là một nguồn của giá trị kinh tế và cách tân mới. Thậm chí còn hơn nữa. Uy thế của dữ liệu lớn tượng trưng cho ba sự thay đổi trong cách chúng ta phân tích thông tin, làm biến đổi cách chúng ta hiểu và tổ chức xã hội.

Sự thay đổi thứ nhất được mô tả trong Chương Hai. Trong thế giới mới này, chúng ta có thể phân tích nhiều dữ liệu hơn hẳn. Trong một số trường hợp, chúng ta thậm chí có thể xử lý tất cả dữ liệu liên quan đến một hiện tượng đặc biệt. Từ thế kỷ thứ mười chín, xã hội đã phụ thuộc vào việc sử dụng các hình mẫu khi phải đối mặt với những số lượng lớn. Tuy nhiên, sự cần thiết phải lấy mẫu là một tạo tác của thời kỳ khan hiếm thông tin, một sản phẩm của những hạn chế tự nhiên khi tương tác với thông tin trong thời đại analog. Trước khi công nghệ kỹ thuật số có hiệu suất cao thịnh hành, chúng ta không hề nhận ra chọn mẫu là những xiềng xích nhân tạo - chúng ta thường hiển nhiên chấp nhận nó. Việc sử dụng tất cả các dữ liệu cho phép chúng ta xem xét những chi tiết chưa hề xem được khi bị giới hạn với những số lượng nhỏ hơn. Dữ liệu lớn cho chúng ta một cái nhìn đặc biệt rõ

ràng về các tiểu phần: tiểu thể loại và tiểu thị trường mà mẫu không thể ước định được.

Việc xem xét dữ liệu rộng lớn hơn cũng cho phép chúng ta nới lỏng mong muốn hướng tới tính chính xác, là sự thay đổi thứ hai, được đề cập tới trong Chương Ba. Đó là một sự đánh đổi: với ít lỗi hơn từ chọn mẫu, chúng ta có thể chấp nhận nhiều lỗi đo lường hơn. Khi khả năng để đo lường là có hạn, chúng ta chỉ tính đến những thứ quan trọng nhất. Sự cố gắng để có được con số chính xác là hợp lý.

Ta không thể bán được gia súc nếu người mua không biết chắc liệu có 100 hay chỉ có 80 con trong đàn. Cho đến gần đây, tất cả các công cụ kỹ thuật số của chúng ta có tiền đề là sự chính xác: chúng ta giả định rằng công cụ cơ sở dữ liệu sẽ truy tìm được các bản ghi hoàn toàn phù hợp với câu hỏi của chúng ta, giống như các bảng tính điện tử lập biểu các con số trong một cột.

Loại tư duy này là một chức năng của môi trường "dữ liệu nhỏ": với rất ít thứ để đo lường, chúng ta phải xem xét những gì quan tâm để định lượng một cách càng chính xác càng tốt.

Theo một số cách nào đó thì việc này là hiển nhiên: một cửa hàng nhỏ có thể đếm tiền trong quỹ cuối ngày tới tận đồng xu, nhưng chúng ta sẽ không - thực sự là không thể - làm tương tự cho tổng sản phẩm nội địa của một quốc gia. Khi quy mô tăng, số lượng của những sự không chính xác cũng tăng.

Tính chính xác đòi hỏi dữ liệu được giám tuyển một cách cẩn thận. Điều này có thể làm được cho những số lượng nhỏ, và tất nhiên một số trường hợp vẫn đòi hỏi như vậy: ta hoặc có hoặc không có đủ tiền trong ngân hàng để viết một chi phiếu. Nhưng đổi lại, khi sử dụng những bộ dữ liệu toàn diện hơn nhiều, chúng ta có thể bỏ đi tính chính xác cứng nhắc trong một thế giới dữ liệu lớn.

Thông thường, dữ liệu lớn là lộn xộn, khác nhau về chất lượng, và được phân bổ giữa vô số các máy chủ trên khắp thế giới. Với dữ liệu lớn, chúng ta sẽ thường hài lòng với khả năng định hướng chung chứ không phải là hiểu biết một hiện tượng chi tiết tới tận xăng-ti-mét, đồng xu, hay nguyên tử. Chúng ta không bỏ qua hoàn toàn sự chính xác; chúng ta chỉ bỏ qua sự sùng bái nó. Những gì chúng ta mất về độ chính xác ở cấp vi mô sẽ được bù đắp lại nhờ cái nhìn sâu sắc ở cấp vĩ mô.

Hai sự thay đổi này dẫn đến một sự thay đổi thứ ba, mà chúng ta giải thích trong Chương Bốn: sự chuyển hướng khỏi việc tìm kiếm lâu đời cho quan hệ nhân quả. Là con người, chúng ta đã được định vị để đi tìm kiếm các nguyên nhân, mặc dù việc tìm kiếm quan hệ nhân quả thường rất khó khăn và có thể dẫn chúng ta lạc đường. Trong một thế giới dữ liệu lớn, ngược lại, chúng ta sẽ không phải gắn chặt vào quan hệ nhân quả; thay vào đó chúng ta có thể khám phá các khuôn mẫu và mối tương quan trong các dữ liệu để thu được những hiểu biết mới lạ và vô giá. Các mối tương quan có thể không cho chúng ta biết chính xác *tại sao* một cái gì đó đang xảy ra, nhưng chúng cảnh báo chúng ta *rằng* cái đó đang xảy ra.

Và trong nhiều tình huống thì điều này là đủ tốt. Nếu hàng triệu hồ sơ y tế điện tử cho thấy những bệnh nhân ung thư nếu dùng một kết hợp nào đó của aspirin và nước cam thì thấy bệnh của họ thuyên giảm, thì nguyên nhân chính xác cho việc cải thiện sức khỏe có thể ít quan trọng hơn so với thực tế là họ sống. Tương tự như vậy, nếu chúng ta có thể tiết kiệm được tiền bằng cách biết thời gian tốt nhất để mua một vé máy bay mà không hiểu các phương pháp phía sau sự điên rồ của vé máy bay, như vậy cũng đủ tốt rồi. Dữ liệu lớn là về *cái gì*, chứ không về *tại sao*. Chúng ta không luôn luôn cần biết nguyên nhân của một hiện tượng, thay vào đó, chúng ta có thể để cho dữ liệu tự nói.

Trước thời dữ liệu lớn, phân tích của chúng ta thường được giới

hạn vào việc thử nghiệm một số lượng nhỏ những giả thuyết được xác định rõ ràng trước khi thu thập dữ liệu. Khi để cho các dữ liệu lên tiếng, chúng ta có thể tạo nên những kết nối mà ta chưa bao giờ nghĩ là chúng tồn tại. Do đó, một số quỹ đầu tư phân tích Twitter để dự đoán hiệu suất của thị trường chứng khoán. Amazon và Netilix căn cứ đề xuất sản phẩm của họ trên vô số các tương tác của người dùng trên các trang này. Twitter, Linkedln và Facebook cũng đều quy chiếu “đồ thị xã hội” các mối quan hệ của người sử dụng để tìm hiểu các sở thích của họ.

Tất nhiên, con người đã phân tích dữ liệu hàng thiên niên kỷ nay. Chữ viết đã được phát triển ở vùng Lưỡng Hà cổ đại bởi các quan chức muốn có một công cụ hiệu quả để ghi lại và theo dõi thông tin. Từ thời Kinh Thánh, các chính phủ đã tổ chức các cuộc điều tra để thu thập các bộ dữ liệu lớn về công dân của họ, và tương tự đã hai trăm năm nay, các chuyên gia tính toán thu thập khối lượng lớn dữ liệu liên quan đến các rủi ro mà họ hy vọng sẽ hiểu được - hoặc ít nhất là tránh được.

Tuy nhiên, trong thời đại analog, việc thu thập và phân tích dữ liệu như vậy là vô cùng tốn kém và mất thời gian. Những câu hỏi mới thường có nghĩa là dữ liệu phải được thu thập lại và việc phân tích phải bắt đầu lại. Bước tiến lớn đối với việc quản lý dữ liệu hiệu quả hơn đã xuất hiện cùng với số hóa: giúp cho máy tính có thể đọc thông tin analog, mà cũng làm cho nó dễ dàng hơn và rẻ hơn để lưu trữ và xử lý.

Bước phát triển này đã cải thiện hiệu quả đáng kể. Việc thu thập và phân tích thông tin trước đây phải mất hàng năm, nay có thể được thực hiện trong vài ngày hoặc thậm chí ngắn hơn. Nhưng rất ít thứ khác thay đổi. Những người phân tích dữ liệu đã quá thường xuyên bị ngập trong thế giới analog, cho rằng các tập dữ liệu chỉ có những mục đích đơn lẻ mà giá trị của chúng đã được gắn liền. Các tiến trình của chúng ta đã duy trì định kiến này. Dù cũng quan trọng như số hóa đã tạo điều kiện cho việc chuyển

sang dữ liệu lớn, nhưng chỉ sự tồn tại của máy tính đã không làm cho dữ liệu lớn xảy ra.

Tuy chưa có thuật ngữ thật tốt để mô tả những gì đang diễn ra hiện nay, nhưng một thuật ngữ giúp định hình được những thay đổi đó là *dữ liệu hóa (datafication)*, một khái niệm mà chúng ta giới thiệu trong Chương Năm. Nó ám chỉ việc lấy thông tin về tất cả mọi thứ dưới ánh mặt trời - bao gồm cả những thứ chúng ta không bao giờ xem là thông tin, chẳng hạn như vị trí của một người, những rung động của một động cơ, hoặc sự căng trên một cây cầu - và biến nó thành một định dạng dữ liệu để thực hiện định lượng nó. Điều này cho phép chúng ta sử dụng thông tin theo những cách mới, chẳng hạn như trong phân tích tiên đoán: phát hiện một động cơ dễ bị sự cố dựa trên độ nóng hay những rung động mà nó tạo ra. Kết quả là chúng ta có thể mở khóa những giá trị tiềm ẩn, bên trong của thông tin.

Có một cuộc truy lùng kho báu đang xảy ra, được thúc đẩy bởi những hiểu biết sâu sắc từ các dữ liệu và giá trị tiềm tàng có thể được khai thông nhờ sự chuyển dịch từ quan hệ nhân quả sang tương liên. Nhưng nó không chỉ là một kho báu. Mỗi bộ dữ liệu riêng lẻ rất có thể có một số giá trị nào đó nội tại, ẩn, chưa được khai phá, và cuộc đua ở đây là để khám phá và nắm bắt tất cả những thứ đó.

Dữ liệu lớn thay đổi bản chất của kinh doanh, thị trường, và xã hội, như chúng ta mô tả trong Chương Sáu và Bảy. Trong thế kỷ XX, giá trị đã chuyển từ cơ sở hạ tầng vật lý như đất đai và nhà máy sang những thứ vô hình như thương hiệu và sở hữu trí tuệ. Điều này bây giờ mở rộng tới dữ liệu, cái đang trở thành một tài sản đáng kể của công ty, một đầu vào kinh tế quan trọng, và là nền tảng của các mô hình kinh doanh mới. Nó là dầu hỏa của nền kinh tế thông tin. Mặc dù dữ liệu hiếm khi được ghi nhận vào bảng cân đối của doanh nghiệp, nhưng điều này có lẽ chỉ là vấn đề thời gian.

Mặc dù một số kỹ thuật nghiền (crunching) dữ liệu đã xuất hiện được một thời gian, trong quá khứ chúng chỉ được dành cho cơ quan tình báo, các phòng nghiên cứu, và các công ty lớn nhất thế giới. Xét cho cùng, Walmart và Capital One đã đi tiên phong trong việc sử dụng dữ liệu lớn trong bán lẻ và ngân hàng, và qua đó làm thay đổi ngành công nghiệp của họ. Bây giờ nhiều trong số những công cụ này đã được dân chủ hóa (mặc dù dữ liệu thì không).

Ảnh hưởng lên các cá nhân có thể là cú sốc lớn nhất. Kinh nghiệm chuyên môn về lĩnh vực đặc thù trở thành ít quan trọng hơn trong một thế giới mà ở đó xác suất và mối tương quan là tối cao. Trong bộ phim *Moneyball,* các tuyển trạch viên bóng chày đã bị các nhà thống kê lấn lướt, khi bản năng nhường chỗ cho các phân tích tinh vi. Tương tự như vậy, các chuyên gia sẽ không biến mất, nhưng họ sẽ phải đối mặt với những điều mà các phân tích dữ liệu lớn thể hiện. Điều này sẽ bắt buộc có sự điều chỉnh những ý tưởng truyền thống của quản lý, ra quyết định, nguồn nhân lực và giáo dục.

Hầu hết các thể chế của chúng ta được thiết lập theo giả định rằng các quyết định của con người được dựa trên thông tin mang bản chất nhỏ lẻ, chính xác, và nhân quả. Nhưng tình hình thay đổi khi dữ liệu là rất lớn, có thể được xử lý một cách nhanh chóng, và chấp nhận sự không chính xác. Hơn nữa, do kích thước rất lớn của dữ liệu, các quyết định có thể thường được thực hiện không bởi con người mà bởi máy. Chúng ta sẽ xem xét những mặt tối của dữ liệu lớn trong Chương Tám. Xã hội đã có hàng thiên niên kỷ trải nghiệm trong việc tìm hiểu và giám sát hành vi của con người. Nhưng làm thế nào để bạn chỉnh đốn một thuật toán? Buổi đầu của tính toán, các nhà hoạch định chính sách công nhận công nghệ có thể được sử dụng để làm suy giảm sự riêng tư ra sao. Kể từ đó xã hội đã xây dựng nhiều quy tắc để bảo vệ thông tin cá nhân. Nhưng trong thời đại của dữ liệu lớn, những luật lệ này tạo thành một dạng Phòng tuyến Maginot gần

như vô dụng. Người ta sẵn sàng chia sẻ thông tin trực tuyến - một tính năng trung tâm của các dịch vụ, không phải là một lỗ hổng để ngăn chặn.

Trong khi đó, mối nguy hiểm đối với những cá nhân như chúng ta chuyển từ yếu tố riêng tư sang xác suất: các thuật toán sẽ dự đoán khả năng một người bị nhồi máu cơ tim (và phải trả nhiều hơn cho bảo hiểm y tế), khả năng vỡ nợ của một khoản thế chấp (và bị từ chối một khoản vay), hoặc phạm tội (và có lẽ bị bắt trước). Nó dẫn đến một sự xem xét mang tính đạo đức về vai trò của tự do ý chí đối với sự độc tài của dữ liệu. Liệu có nên để ý chí cá nhân chiến thắng dữ liệu lớn, ngay cả khi số liệu thống kê lý giải khác? Cũng giống như việc in ấn đã chuẩn bị nền tảng cho các đạo luật đảm bảo tự do ngôn luận - điều không tồn tại trước đó bởi có rất ít việc biểu đạt bằng văn bản cần được bảo vệ - thời đại của dữ liệu lớn sẽ đòi hỏi những quy định mới để bảo vệ sự thiêng liêng của cá nhân.

Dù gì đi nữa, cách thức chúng ta kiểm soát và xử lý dữ liệu sẽ phải thay đổi. Chúng ta đang bước vào một thế giới của những dự đoán liên tục dựa trên dữ liệu, ở đó chúng ta có thể không giải thích được các nguyên nhân đằng sau những quyết định của chúng ta. Thử hỏi còn có ý nghĩa gì khi bác sĩ không thể biện minh cho biện pháp can thiệp y tế của mình nếu không yêu cầu bệnh nhân trông chờ vào một cái hộp đen, giống như bác sĩ phải làm khi dựa vào chẩn đoán được dẫn dắt bởi dữ liệu lớn? Liệu chuẩn mực "chứng cứ hợp lý" của hệ thống tư pháp có cần phải thay đổi thành "chứng cứ theo xác suất" - và nếu như vậy thì hệ quả của điều này là những gì đối với tự do và phẩm giá con người?

Những nguyên tắc mới là cần thiết cho thời đại của dữ liệu lớn, mà chúng ta đặt ra trong Chương Chín. Mặc dù chúng được xây dựng dựa trên các giá trị đã được phát triển và được ghi nhận đối với thế giới của dữ liệu nhỏ, điều đó không đơn giản là vấn đề

làm mới lại những quy định cũ cho hoàn cảnh mới, mà là hoàn toàn công nhận sự cần thiết của những nguyên tắc mới.

Những lợi ích cho xã hội sẽ là vô kể, khi dữ liệu lớn trở thành bộ phận của giải pháp cho những vấn đề bức xúc toàn cầu, như giải quyết thay đổi khí hậu, xóa bỏ bệnh tật, thúc đẩy sự quản trị tốt và phát triển kinh tế. Nhưng thời đại dữ liệu lớn cũng thách thức chúng ta phải chuẩn bị tốt hơn về những cách thức trong đó việc khai thác công nghệ sẽ làm thay đổi các tổ chức của chúng ta và chính bản thân chúng ta.

Dữ liệu lớn đánh dấu một bước quan trọng trong việc tìm kiếm của con người để định lượng và hiểu thế giới; một ưu thế của những thứ chưa bao giờ được đo lường, lưu trữ, phân tích và chia sẻ trước khi được dữ liệu hóa. Việc khai thác lượng lớn dữ liệu thay vì chỉ một phần nhỏ, và việc có đặc quyền với nhiều dữ liệu có độ chính xác thấp hơn, sẽ mở ra cánh cửa tới những cách hiểu biết mới. Nó dẫn xã hội tới việc từ bỏ ưu tiên lâu đời cho nhân quả, và trong nhiều trường hợp thu được các lợi ích của mối tương liên.

Lý tưởng về việc xác định được những cơ chế nhân-quả chỉ là một kiểu ảo tưởng tự mãn; dữ liệu lớn đã làm đảo lộn điều này. Một lần nữa chúng ta đang lâm vào một sự bế tắc lịch sử nơi "thần thánh cũng chết", nghĩa là những điều chắc chắn chúng ta đã từng tin vào, một lần nữa lại thay đổi. Nhưng lần này chúng được thay thế một cách thật trớ trêu bằng những chứng cứ tốt hơn. Vậy thì trực giác, niềm tin, và những điều mơ hồ sẽ còn lại vai trò gì, so với các chứng cứ và việc học tập bằng trải nghiệm? Khi thế giới chuyển từ quan hệ nhân quả sang tương liên, làm sao chúng ta có thể tiến một cách thực dụng về phía trước mà không làm suy yếu những nền tảng của xã hội, nhân loại, và tiến bộ dựa trên nhân-quả?

Cuốn sách này mong muốn giải thích chúng ta đang ở đâu, dõi

theo dấu vết chúng ta đã tới đây như thế nào, và cung cấp một hướng dẫn hết sức cần thiết về những lợi ích và những nguy hiểm nằm ở phía trước.

2. NHIỀU HƠN

DỮ LIỆU LỚN ĐỀU LIÊN QUAN ĐẾN về việc nhìn và hiểu các mối quan hệ trong và giữa các mẩu thông tin, mà cho đến rất gần đây, chúng ta phải chật vật để nắm bắt được một cách đầy đủ. Theo chuyên gia dữ-liệu-lớn của IBM Jeff Jonas, bạn cần để cho dữ liệu “nói với mình”. Ở một mức độ nào đó điều này nghe có vẻ hiển nhiên. Con người đã xem xét dữ liệu để tìm hiểu về thế giới trong một thời gian dài, cho dù theo nghĩa không chính thức của vô số các quan sát chúng ta thực hiện mỗi ngày, chủ yếu là trong vài thế kỷ vừa qua, hay theo ý nghĩa chính thức của các đơn vị định lượng có thể được xử lý bằng những thuật toán mạnh mẽ.

Thời đại kỹ thuật số có thể đã làm cho việc xử lý dữ liệu dễ dàng hơn và nhanh hơn, để tính toán hàng triệu con số chỉ trong tích tắc. Nhưng khi đề cập đến việc dữ liệu lên tiếng, chúng ta đề cập tới một điều gì đó nhiều hơn - và khác hơn. Như đã lưu ý trong Chương Một, dữ liệu lớn là về ba sự chuyển đổi lớn lao của tư duy được nối kết với nhau và do đó củng cố lẫn nhau. Thứ nhất là khả năng phân tích lượng lớn dữ liệu về một chủ đề thay vì bị buộc phải thỏa mãn với những tập hợp nhỏ hơn. Thứ hai là sự sẵn sàng để đón nhận sự hỗn độn trong thế giới thực của dữ liệu thay vì đòi hỏi đặc quyền về tính chính xác. Thứ ba là sự tôn trọng ngày càng tăng đối với các mối tương quan thay vì việc tiếp tục truy tìm nhân quả rất khó nắm bắt. Chương này xem xét sự thay đổi thứ nhất: sử dụng tất cả các dữ liệu ta có thay vì chỉ một phần nhỏ của nó.

Thách thức trong việc xử lý những khối lượng lớn dữ liệu thực chất đã tồn tại từ khá lâu. Trong gần hết lịch sử, chúng ta đã làm việc với chỉ một ít dữ liệu vì các công cụ để thu thập, tổ chức, lưu trữ, và phân tích nó rất nghèo nàn. Chúng ta sàng lọc thông tin, giữ lại mức tối thiểu vừa đủ để có thể khảo sát được dễ dàng hơn. Đây là một hình thức của tự kiểm duyệt vô thức: chúng ta xử lý các khó khăn trong việc tương tác với dữ liệu như thể đó là

những chuyện không may, chứ không phải như bản chất thật của nó - một hạn chế nhân tạo bị áp đặt bởi công nghệ vào thời điểm đó. Ngày nay, môi trường kỹ thuật đã thay đổi 179 độ. Vẫn còn, và luôn luôn sẽ còn, một hạn chế về dung lượng dữ liệu chúng ta có thể quản lý, nhưng hạn chế đó là ít hơn nhiều so với trước đây và sẽ càng ít hơn trong tương lai.

Theo một số cách nào đó, chúng ta vẫn chưa hoàn toàn đánh giá cao sự tự do mới của mình trong việc thu thập và sử dụng những khối lớn dữ liệu. Hầu hết kinh nghiệm và thiết kế tổ chức của chúng ta đã giả định rằng sự sẵn có của thông tin là hạn chế. Chúng ta chấp nhận chỉ có thể thu thập được một ít thông tin, và đó thường là những gì chúng ta đã làm. Nó đã trở thành sự tự thỏa mãn.

Chúng ta thậm chí còn phát triển các kỹ thuật phức tạp để sử dụng ít dữ liệu nhất có thể. Xét cho cùng, một mục đích của thống kê là để xác nhận một điều khám phá tuyệt vời nhất bằng cách sử dụng lượng dữ liệu ít nhất. Trong thực tế, chúng ta đã hệ thống hóa việc thực thi của mình để bóp nghẹt lượng thông tin chúng ta sử dụng trong các định mức, tiến trình, và cơ chế khuyến khích. Để có được một sự hình dung về ý nghĩa của sự chuyển dịch tới dữ liệu lớn, câu chuyện bắt đầu với một cái nhìn ngược thời gian.

Cho đến gần đây các công ty tư nhân, và ngày nay ngay cả các cá nhân, đã có thể thu thập và sắp xếp thông tin trên một quy mô lớn. Trước đây, công việc này thuộc các tổ chức lớn hơn như nhà thờ và nhà nước, mà trong nhiều xã hội chúng là đồng nhất. Ghi nhận lâu đời nhất của việc đếm là từ khoảng 5000 năm trước công nguyên, khi các thương nhân Sumer sử dụng những cục đất sét nhỏ để biểu thị hàng hóa khi buôn bán. Tuy nhiên việc đếm trên một quy mô lớn hơn lại thuộc phạm vi hoạt động của nhà nước. Qua nhiều thiên niên kỷ, các chính phủ đã cố gắng kiểm soát người dân của họ bằng cách thu thập thông tin.

Hãy xem việc điều tra dân số. Người Ai Cập cổ đại được cho là đã tiến hành những cuộc điều tra dân số, cũng như người Trung Hoa. Những việc này được đề cập đến trong Cựu Ước, và Tân Ước cho chúng ta biết Caesar Augustus đã áp đặt một cuộc điều tra dân số - "cả thế giới nên bị đánh thuế" - đưa Joseph và Maria đến Bethlehem, nơi Jesus đã sinh ra. Cuốn *Domesday Book* năm 1086, một trong những báu vật được sùng kính nhất của người Anh, tại thời gian đó, là một sự kiểm đếm toàn diện chưa từng có về người Anh cùng đất đai và tài sản của họ. Các ủy viên hoàng gia đã di khắp nơi, tổng hợp thông tin để đưa vào cuốn sách - sau đó mới có tên *Domesday,* hoặc *Khải huyền*, bởi vì quá trình này giống như Phán xét cuối cùng trong Kinh Thánh, khi cuộc sống của tất cả mọi người bị phơi bày.

Tiến hành điều tra dân số luôn tốn tiền và tốn thời gian. Vua William I, người ra lệnh thực hiện *Domesday Book*, đã không còn sống để nhìn thấy nó được hoàn thành. Nhưng lựa chọn duy nhất để khỏi phải mang gánh nặng này là từ bỏ thu thập thông tin. Và ngay cả sau khi tốn tất cả thời gian và chi phí, thông tin vẫn chỉ là gần đúng, vì những người đi điều tra không thể đếm được tất cả mọi người một cách hoàn hảo. Từ "điều tra dân số" xuất phát từ thuật ngữ La-tinh "censere" có nghĩa là "để ước tính". Hơn ba trăm năm trước, một người Anh bán đồ may vá tên John Graunt đã có một ý tưởng mới lạ. Graunt muốn biết dân số London tại thời điểm bệnh dịch hạch. Thay vì đếm mỗi người, ông đã nghĩ ra một cách tiếp cận - mà ngày nay chúng ta gọi là "thống kê" - cho phép ông *suy* ra quy mô dân số. Cách tiếp cận của ông là thô, nhưng nó thiết lập ý tưởng rằng người ta có thể ngoại suy từ một mẫu nhỏ những hiểu biết hữu ích về dân số tổng quát. Nhưng cách người ta làm thế nào mới quan trọng. Graunt thì chỉ nhân rộng ra từ mẫu của mình.

Hệ thống của ông đã nổi tiếng, mặc dù sau đó chúng ta biết những con số của ông là hợp lý nhờ may mắn. Trải qua nhiều thế hệ, việc chọn mẫu vẫn sai sót rất lớn. Do đó với các cuộc điều tra

dân số và những công việc dạng “dữ liệu lớn” tương tự, cách tiếp cận để cố gắng đếm tất cả vẫn là phổ biến.

Bởi các cuộc điều tra dân số rất phức tạp, tốn chi phí và tốn thời gian, nên chúng ít được thực hiện. Người La Mã cổ đại, vẫn tự hào với một dân số mấy trăm ngàn, thực hiện điều tra dân số năm năm một lần. Hiến pháp Hoa Kỳ bắt buộc một cuộc điều tra dân số trong mỗi thập kỷ, khi đất nước đang phát triển này có tới hàng triệu người. Nhưng vào cuối thế kỷ XIX, thậm chí việc này cũng trở nên khó khăn. Dữ liệu đã vượt quá khả năng xử lý của Cục Điều tra Dân số.

Điều gây sốc là cuộc điều tra dân số năm 1880 đã mất tám năm để hoàn thành. Thông tin đã trở thành lỗi thời ngay cả trước khi nó được công bố. Tệ hơn nữa, các quan chức ước tính việc điều tra dân số năm 1890 sẽ cần tới 13 năm để lập bảng - một tình trạng hết sức vô lý, chưa nói đến chuyện vi phạm Hiến pháp. Tuy nhiên, do việc phân chia các loại thuế và đại diện trong Quốc hội dựa trên dân số, nên việc có được không chỉ một con số chính xác mà còn phải kịp thời là rất cần thiết.

Vấn đề Cục Điều tra Dân số Hoa Kỳ phải đối mặt cũng tương tự với sự khó khăn của các nhà khoa học và doanh nhân vào đầu thiên niên kỷ mới, khi vấn đề trở nên rõ ràng là họ đã chết đuối trong dữ liệu: số lượng thông tin được thu thập đã hoàn toàn tràn ngập các công cụ được sử dụng để xử lý chúng, và người ta bắt buộc cần tới những kỹ thuật mới. Trong những năm 1880 tình hình nghiêm trọng tới mức Cục Điều tra Dân số ký hợp đồng với Herman Hollerith, một nhà phát minh người Mỹ, để sử dụng ý tưởng của ông về thẻ đục lỗ và máy lập bảng cho điều tra dân số năm 1890.

Với nỗ lực rất lớn, ông đã thành công trong việc rút ngắn thời gian lập bảng từ tám năm xuống dưới một năm. Đó là một thành tích tuyệt vời, đánh dấu việc bắt đầu xử lý dữ liệu tự động (và

cung cấp nền tảng cho những gì sau này trở thành IBM).

Nhưng như một phương pháp thu nhận và phân tích dữ liệu lớn, nó vẫn còn rất tốn kém. Rốt cuộc, mỗi người tại Hoa Kỳ đều phải điền vào một mẫu đơn và các thông tin được chuyển vào một thẻ đục lỗ, được sử dụng để lập bảng. Với các phương pháp tốn kém như vậy, thật khó tưởng tượng nổi có thể thực hiện một cuộc điều tra dân số trong bất kỳ khoảng thời gian nào ngắn hơn một thập kỷ, mặc dù sự chậm trễ là không có ích lợi cho một quốc gia đang phát triển nhảy vọt.

Vấn đề là ở chỗ: Sử dụng tất cả dữ liệu, hay chỉ một chút ít? Lấy tất cả dữ liệu về những gì đang được đo đạc chắc chắn là điều hợp lý nhất. Nó chỉ không phải lúc nào cũng thực tế khi quy mô là rất lớn. Nhưng làm thế nào để chọn một mẫu? Một số người cho rằng việc xây dựng có mục đích một mẫu đại diện được cho toàn bộ sẽ là cách phù hợp nhất. Nhưng vào năm 1934, Jerzy Neyman, một nhà thống kê Ba Lan, đã chứng minh một cách ấn tượng rằng cách tiếp cận như vậy dẫn đến những sai sót rất lớn. Chìa khóa để tránh chúng là nhằm vào sự ngẫu nhiên để chọn thành phần đưa vào mẫu.

Các nhà thống kê đã chỉ ra rằng độ chính xác chọn mẫu được cải thiện rất đáng kể với sự ngẫu nhiên, chứ không phải với việc gia tăng kích thước mẫu. Trên thực tế, mặc dù nó có vẻ lạ thường, một mẫu được chọn ngẫu nhiên của 1.100 quan sát riêng lẻ trên một câu hỏi nhị phân (có hay không, với khoảng tỷ lệ bằng nhau) là đại diện đáng kể cho toàn dân. 19 trong 20 trường hợp, nó nằm trong khoảng biên độ 3 phần trăm lỗi, bất kể quy mô tổng dân số là một trăm ngàn hay một trăm triệu người. Lý do của điều này lại rất phức tạp về mặt toán học, nhưng câu trả lời ngắn gọn là sau một điểm nhất định, khi các con số ngày càng lớn lên, thì số lượng biên của thông tin mới mà chúng ta thu được từ mỗi quan sát sẽ ngày càng nhỏ đi. Thực tế, sự ngẫu nhiên quan trọng hơn cỡ mẫu là một hiểu biết sâu sắc đáng ngạc nhiên. Nó đã mở

đường cho một cách tiếp cận mới để thu thập thông tin.

Dữ liệu sử dụng các mẫu ngẫu nhiên có thể được thu thập với chi phí thấp nhưng được ngoại suy với độ chính xác cao cho tổng thể. Kết quả là các chính phủ có thể tiến hành các phiên bản nhỏ của tổng điều tra sử dụng các mẫu ngẫu nhiên mỗi năm, thay vì chỉ làm một tổng điều tra trong mỗi thập kỷ. Và họ đã làm như vậy. Ví dụ Cục Điều tra Dân số Hoa Kỳ thực hiện hơn 200 cuộc điều tra kinh tế và dân số hàng năm dựa trên cơ sở lấy mẫu, để bổ sung cho cuộc tổng điều tra dân số mười năm một lần trong đó cố gắng đếm tất cả mọi người. Lấy mẫu là một giải pháp cho vấn đề quá tải thông tin trước đây, khi việc thu thập và phân tích dữ liệu rất khó thực hiện.

Các ứng dụng của phương pháp mới này nhanh chóng vượt ra khỏi khu vực công và các cuộc tổng điều tra. Về bản chất, lấy mẫu ngẫu nhiên làm giảm những vấn đề dữ liệu lớn xuống thành những vấn đề dữ liệu dễ quản lý hơn. Trong kinh doanh, nó được sử dụng để đảm bảo chất lượng sản xuất - làm cho các cải tiến trở nên dễ dàng hơn và ít tốn kém hơn. Kiểm tra chất lượng toàn diện lúc đầu đòi hỏi phải nhìn vào từng sản phẩm đơn lẻ đi ra từ băng chuyền; bây giờ một mẫu ngẫu nhiên để kiểm tra cho một loạt sản phẩm là đủ. Tương tự như vậy, phương pháp mới đã mở ra các cuộc khảo sát người tiêu dùng trong bán lẻ và các cuộc thăm dò trong chính trị. Nó đã chuyển đổi một phần đáng kể những gì chúng ta vẫn gọi là các ngành nhân văn trở thành các ngành *khoa học* xã hội.

Lấy mẫu ngẫu nhiên đã là một thành công lớn và là xương sống của đo lường hiện đại có quy mô lớn. Nhưng nó chỉ là một đường tắt, một lựa chọn tốt thứ hai để thu thập và phân tích tập dữ liệu đầy đủ. Nó đi kèm với một số điểm yếu cố hữu. Độ chính xác của nó phụ thuộc vào việc đảm bảo tính ngẫu nhiên khi thu thập dữ liệu mẫu, nhưng đạt được ngẫu nhiên như vậy là khó khăn. Những thành kiến có hệ thống trong cách thức dữ liệu được thu

thập có thể dẫn đến các kết quả ngoại suy rất sai.

Có những dẫn chứng cho những vấn đề như vậy trong phỏng vấn bầu cử sử dụng điện thoại cố định. Mẫu bị thành kiến đối với những người chỉ sử dụng điện thoại di động (những người trẻ hơn và tự do hơn), như nhà thống kê Nate Silver đã chỉ ra. Điều này đã dẫn đến những dự đoán bầu cử không chính xác. Trong cuộc bầu cử tổng thống năm 2008 giữa Barack Obama và John McCain, các tổ chức thăm dò chính của Gallup, Pew, và ABC/Washington Post tìm thấy sự khác biệt từ một đến ba điểm phần trăm, khi họ thăm dò có và không có sự điều chỉnh cho người sử dụng điện thoại di động - một biên độ đáng kể nếu xét tới độ sát sao của cuộc đua.

Rắc rối nhất là lấy mẫu ngẫu nhiên không dễ dàng mở rộng được để bao gồm các tiểu thể loại, vì khi chia kết quả thành các nhóm con nhỏ hơn sẽ làm tăng khả năng dự đoán sai. Thật dễ dàng hiểu lý do. Giả sử bạn thăm dò ý kiến một mẫu ngẫu nhiên của 1.000 người về ý định bỏ phiếu của họ trong cuộc bầu cử sắp tới. Nếu mẫu của bạn là đủ ngẫu nhiên, khả năng có thể xảy ra là ý kiến của toàn bộ dân số sẽ ở trong phạm vi 3 phần trăm của các quan điểm trong mẫu. Nhưng sẽ ra sao nếu cộng hoặc trừ 3 phần trăm là không đủ chính xác? Hoặc sẽ ra sao nếu sau đó bạn muốn chia nhóm thành những nhóm nhỏ hơn, với giới tính, địa lý, hoặc thu nhập?

Và điều gì sẽ xảy ra nếu bạn muốn kết hợp các phân nhóm này để nhắm tới một nhóm dân số thích hợp? Trong một mẫu tổng thể của 1.000 người, một phân nhóm như “nữ cử tri giàu có ở vùng Đông Bắc” sẽ nhỏ hơn 100 nhiều. Chỉ sử dụng vài chục quan sát để dự đoán những ý định bỏ phiếu của *tất cả* các nữ cử tri giàu có ở vùng Đông Bắc sẽ là không chính xác ngay cả với sự ngẫu nhiên gần như hoàn hảo. Và những thành kiến nhỏ nhất trong mẫu tổng thể sẽ làm cho các lỗi trở thành rõ rệt hơn ở mức độ phân nhóm.

Do đó, việc lấy mẫu một cách nhanh chóng không còn hữu ích khi bạn muốn đi sâu hơn, để có một cái nhìn gần hơn đối với một số tiểu thể loại hấp dẫn trong dữ liệu. Những gì hoạt động được ở tầm vĩ mô lại thất bại hoàn toàn ở tầm vi mô. Lấy mẫu giống như một bức in ảnh analog. Nó trông đẹp từ một khoảng cách, nhưng khi bạn ngắm gần hơn, phóng to một chi tiết đặc biệt thì nó bị mờ. Lấy mẫu cũng đòi hỏi phải lập kế hoạch và thực hiện cẩn thận. Người ta thường không thể "hỏi" mẫu những câu hỏi mới nếu chúng chưa được dự liệu ngay từ đầu. Vì vậy, mặc dù là một đường tắt rất hữu ích, sự đánh đổi ở đây quả thực chỉ đơn thuần là một đường tắt. Khi là một mẫu chứ không phải tất cả, tập dữ liệu thiếu khả năng mở rộng nhất định hoặc tính mềm dẻo, theo đó cùng một dữ liệu có thể được phân tích lại theo một cách hoàn toàn mới so với mục đích mà ban đầu nó được thu thập.

Hãy xem xét trường hợp phân tích DNA. Chi phí để xác định trình tự gen của một cá nhân là gần 1.000 đôla vào năm 2012, khiến nó gần trở thành một kỹ thuật thị trường đại chúng có thể được thực hiện theo quy mô lớn. Kết quả là một ngành công nghiệp mới xác định trình tự gen cá nhân được ra đời. Từ năm 2007, công ty 23andMe ở Thung Lũng Silicon đã phân tích DNA của người với giá chỉ vài trăm đôla. Kỹ thuật của nó có thể tiết lộ những đặc điểm trong mã di truyền của người có thể làm cho họ dễ bị mắc một số bệnh như ung thư vú hoặc các vấn đề về tim. Và bằng cách tập hợp thông tin DNA và sức khỏe của khách hàng, 23andMe hy vọng sẽ học hỏi được những điều mới mẻ không thể phát hiện được bằng những phương cách khác.

Nhưng có một cản trở. Công ty xác định trình tự chỉ một phần nhỏ mã di truyền của một người: những nơi đã được biết là dấu hiệu cho thấy những điểm yếu di truyền đặc biệt. Trong khi đó, hàng tỷ cặp DNA cơ sở vẫn chưa được xác định trình tự. Do đó 23andMe chỉ có thể trả lời những câu hỏi về các dấu hiệu mà nó xem xét. Bất cứ khi nào một dấu hiệu mới được phát hiện, DNA của một người (hay chính xác hơn, phần liên quan của nó) phải

được xác định trình tự lại. Làm việc với một tập hợp con, chứ không phải là toàn bộ, đòi hỏi một sự đánh đổi: công ty có thể thấy những gì họ tìm kiếm một cách nhanh hơn và rẻ hơn, nhưng nó không thể trả lời được những câu hỏi mà nó không xem xét từ trước.

Giám đốc điều hành huyền thoại Steve Jobs của Apple đã thực hiện một tiếp cận hoàn toàn khác trong cuộc chiến của ông chống lại bệnh ung thư. Ông trở thành một trong những người đầu tiên trên thế giới để toàn bộ DNA của mình cũng như của khối u của ông được xác định trình tự. Để làm điều này, ông đã trả một khoản tiền sáu con số - hàng trăm lần so với giá 23andMe tính. Đổi lại, ông đã nhận được không phải một mẫu, một tập hợp nhỏ các dấu hiệu, mà là một tệp dữ liệu chứa toàn bộ các mã di truyền.

Khi lựa chọn thuốc cho một bệnh nhân ung thư thông thường, các bác sĩ phải hy vọng DNA của bệnh nhân là đủ tương tự như của những người tham gia vào thử nghiệm loại thuốc. Còn đội ngũ bác sĩ của Steve Jobs thì có thể lựa chọn các phương pháp điều trị theo cách chúng tác động tốt như thế nào đối với cấu tạo di truyền cụ thể của ông. Bất cứ khi nào một hướng điều trị mất hiệu quả vì ung thư đột biến và kháng cự được nó, các bác sĩ có thể chuyển sang một loại thuốc khác - “nhảy từ một giỏ hoa huệ này sang một giỏ khác”, như Jobs từng mô tả. “Tôi hoặc sẽ là một trong những người đầu tiên có thể chạy nhanh hơn căn bệnh ung thư như thế này hoặc sẽ là một trong những người cuối cùng chết vì nó”, ông nói đùa. Mặc dù rất đáng buồn khi dự đoán của ông không được hoàn thành, những phương pháp này - có tất cả các dữ liệu, chứ không chỉ một phần nhỏ - đã cho ông thêm nhiều năm sống.

Từ một số tới tất cả

Lấy mẫu là một kết quả tự nhiên trong thời đại của những hạn

chế về xử lý thông tin, khi con người đo đạc thế giới nhưng lại thiếu các công cụ để phân tích những gì họ thu thập được.

Thế nên nó cũng là một di tích của thời đại ấy. Những khiếm khuyết trong tính toán và lập bảng hiện nay không còn tồn tại ở cùng mức độ đó nữa. Các cảm biến, điện thoại di động GPS, những cú nhấp chuột trên web, và Twitter thu thập dữ liệu thụ động; máy tính có thể nghiền các con số này ngày càng dễ dàng hơn.

Tuy nhiên, việc lấy mẫu đi kèm với một chi phí mà từ lâu đã được thừa nhận nhưng bị đẩy sang một bên: Nó làm mất đi chi tiết. Trong một số trường hợp, rõ ràng không có cách nào khác ngoài lấy mẫu. Tuy nhiên, trong nhiều lĩnh vực đang diễn ra một sự thay đổi từ thu thập một số dữ liệu sang thu thập càng nhiều càng tốt, và nếu có thể, thì lấy tất cả mọi thứ: *N = tất cả.*

Như chúng ta đã thấy, sử dụng N = tất cả có nghĩa chúng ta có thể đi sâu vào dữ liệu; mẫu không thể làm được điều đó. Thứ hai, hãy nhớ lại rằng trong ví dụ về lấy mẫu ở trên, chúng ta chỉ có một biên độ 3 phần trăm lỗi khi ngoại suy cho toàn bộ dân số. Đối với một số tình huống, biên độ lỗi đó là tốt. Nhưng bạn bị mất các chi tiết, độ chi tiết, khả năng xem xét kỹ hơn ở những phân nhóm nhất định. Một phân phối chuẩn, than ôi, chỉ đạt mức tiêu chuẩn. Thông thường, những điều thực sự thú vị trong cuộc sống lại được tìm thấy ở những nơi mà mẫu không nắm bắt được đầy đủ.

Do đó Xu hướng Dịch cúm của Google không dựa trên một mẫu ngẫu nhiên nhỏ, mà thay vào đó sử dụng hàng tỷ truy vấn Internet ở Mỹ. Việc sử dụng tất cả .dữ liệu chứ không phải chỉ một mẫu nhỏ đã cải thiện việc phân tích sâu xuống tới mức dự đoán được sự lây lan của bệnh cúm trong một thành phố cụ thể chứ không phải chỉ trong một tiểu bang hay toàn bộ quốc gia.

Oren Etzioni của Farecast ban đầu đã sử dụng 12 ngàn điểm dữ liệu, một mẫu, và nó đã hoạt động tốt.

Nhưng khi Etzioni thêm nhiều dữ liệu hơn, chất lượng của các dự báo được cải thiện. Cuối cùng, Farecast đã sử dụng các hồ sơ chuyến bay nội địa của hầu hết các tuyến đường trong cả một năm. “Đây là dữ liệu tạm thời - bạn chỉ cần tiếp tục thu thập nó theo thời gian, và khi bạn làm như vậy, bạn sẽ có được cái nhìn ngày càng sâu sắc hơn vào các khuôn mẫu”, Etzioni cho biết.

Vì vậy, chúng ta sẽ luôn thấy ổn khi bỏ con đường tắt lấy mẫu ngẫu nhiên sang bên và nhắm tới dữ liệu toàn diện hơn để thay thế. Làm như vậy đòi hỏi phải có sức mạnh xử lý và lưu trữ phong phú cũng như các công cụ tiên tiến để phân tích tất cả. Nó cũng đòi hỏi những cách thức dễ dàng và giá cả phải chăng để thu thập dữ liệu. Trong quá khứ, mỗi thứ này là một thách đố đắt giá. Nhưng hiện nay chi phí và độ phức tạp của tất cả các mảnh ghép này đã giảm đáng kể. Những gì trước đây là phạm vi của chỉ các công ty lớn nhất thì bây giờ lại khả thi cho hầu như tất cả.

Sử dụng tất cả dữ liệu cho phép phát hiện các kết nối và chi tiết mà bình thường sẽ bị che giấu trong sự bao la của thông tin. Ví dụ, việc phát hiện các gian lận thẻ tín dụng hoạt động bằng cách tìm kiếm những bất thường, và cách tốt nhất để tìm ra chúng là nghiền tất cả dữ liệu thay vì một mẫu. Các giá trị ngoại lai là những thông tin thú vị nhất, và bạn chỉ có thể nhận ra chúng khi so sánh với hàng loạt giao dịch bình thường. Nó là một vấn đề về dữ liệu lớn. Và bởi vì các giao dịch thẻ tín dụng xảy ra tức thời, nên việc phân tích thường phải được thực hiện trong thời gian thực.

Xoom là một công ty chuyên về chuyển tiền quốc tế và được hỗ trợ bởi những tên tuổi lớn trong lĩnh vực dữ liệu lớn. Nó phân tích tất cả dữ liệu liên quan tới các giao dịch mà nó xử lý. Hệ

thống tăng mức báo động vào năm 2011 khi nó nhận thấy số lượng giao dịch thẻ Discovery có nguồn gốc từ New Jersey hơi cao hơn một chút so với trung bình. “Nó nhận thấy một mô hình mà đáng ra không được như vậy”, John Kunze, giám đốc điều hành của Xoom, giải thích. Xét riêng thì mỗi giao dịch có vẻ hợp pháp. Nhưng cuối cùng thì hóa ra chúng đến từ một nhóm tội phạm. Cách duy nhất để phát hiện sự bất thường là khảo sát tất cả dữ liệu - việc lấy mẫu có thể đã bỏ sót nó.

Sử dụng tất cả các dữ liệu không nhất thiết phải là một công việc rất lớn. Dữ liệu lớn không cần thiết phải lớn một cách tuyệt đối, mặc dù thường thì nó là như vậy. Xu hướng Dịch cúm của Google điều chỉnh các dự đoán của nó trên hàng trăm triệu bài tập mô hình hóa toán học sử dụng hàng tỷ điểm dữ liệu. Việc xác định trình tự đầy đủ của một gen người đưa đến con số ba tỷ cặp cơ sở. Nhưng chỉ xét riêng con số tuyệt đối của các điểm dữ liệu, kích thước của bộ dữ liệu, thì không phải là điều làm cho những thứ này thành những ví dụ của dữ liệu lớn. Thứ xếp loại chúng thành dữ liệu lớn là thay vì sử dụng đường tắt của một mẫu ngẫu nhiên, cả Xu hướng Dịch cúm và các bác sĩ của Steve Jobs đều đã sử dụng toàn bộ dữ liệu ở mức nhiều nhất mà họ có thể.

Phát hiện ra chuyện gian lận trong thi đấu của môn thể thao quốc gia của Nhật Bản, đấu vật sumo, là một minh họa hay tại sao sử dụng N = tất cả không nhất thiết có nghĩa là lớn. Những trận đấu bị dàn xếp vốn luôn bị buộc tội phá hoại môn thể thao của các hoàng đế, và người ta luôn hùng hồn chối biến. Steven Levitt, một nhà kinh tế tại Đại học Chicago, đã xem xét những sai trái trong bộ hồ sơ hơn một thập kỷ của các trận đấu gần đây - tất cả các trận đấu. Trong một bài nghiên cứu thú vị được công bố trên tờ *American Economic Review* và được đăng lại trong cuốn sách *Freakonomics,* ông và một đồng nghiệp đã mô tả tính hữu ích của việc khảo sát nhiều dữ liệu như vậy.

Họ đã phân tích 11 năm số liệu của các trận đấu sumo, hơn

64.000 trận đấu vật, để săn lùng những sự bất thường. Và họ đã bắt được vàng. Việc dàn xếp trận đấu đã thực sự diễn ra, nhưng không phải ở nơi hầu hết mọi người nghi ngờ. Thay vì trong những cuộc đọ sức tranh ngôi vô địch, có thể bị gian lận hoặc không, dữ liệu cho thấy một điều hài hước đã xảy ra trong các trận đấu kết thúc giải vốn không mấy ai chú ý. Có vẻ như ít thứ bị đe dọa, vì các đô vật này không còn cơ hội chiến thắng một danh hiệu nào.

Tuy nhiên, một đặc thù của sumo là đô vật cần phải thắng nhiều hơn thua tại các giải 15 trận đấu để duy trì thứ hạng và thu nhập của họ. Điều này đôi khi dẫn đến sự chênh lệch về lợi ích, khi một đô vật với tỷ lệ 7-7 sẽ gặp một đối thủ có tỉ lệ 8-6 hoặc tốt hơn. Kết quả có ý nghĩa rất lớn đối với đô vật thứ nhất và không có tí ý nghĩa gì cho người thứ hai. Trong những trường hợp này, việc phân tích số liệu đã cho thấy rằng đô vật cần chiến thắng thường sẽ giành chiến thắng.

Những người cần chiến thắng đã thi đấu kiên cường hơn chăng? Có lẽ. Nhưng các dữ liệu cho thấy còn có một cái gì đó khác nữa xảy ra. Các đô vật cần thắng thường thắng khoảng 25 phần trăm nhiều hơn bình thường. Thật khó để gán một sự khác biệt lớn đến vậy cho riêng chỉ hoóc-môn kích thích từ tuyến thượng thận adrenaline. Khi dữ liệu được phân tích xa hơn, nó cho thấy ngay lần kế tiếp hai đô vật gặp lại, người thua trong trận trước rất thường giành chiến thắng so với khi họ thi đấu trong những trận về sau. Vì vậy, chiến thắng đầu tiên dường như là một "món quà" của một đối thủ cạnh tranh cho đối thủ kia, vì đặc điểm có qua có lại trong thế giới đan chen chặt chẽ của sumo.

Thông tin này vẫn luôn luôn rõ ràng. Nó tồn tại sờ sờ trước mắt. Tuy nhiên việc lấy mẫu ngẫu nhiên của các trận đấu đã không tiết lộ nó. Lý do là mặc dù nó dựa trên các thống kê cơ bản, nhưng nếu không biết tìm kiếm cái gì, người ta sẽ không biết phải sử dụng mẫu nào. Ngược lại, Levitt và đồng nghiệp của ông

đã phát hiện ra nó bằng cách sử dụng một tập hợp dữ liệu lớn hơn nhiều - cố gắng kiểm tra toàn bộ các trận đấu. Một cuộc điều tra sử dụng dữ liệu lớn gần giống như một chuyên đi câu: ngay từ đầu nó đã không rõ ràng, kể cả chuyện liệu có câu được món nào chăng và món đó có thể là *cái gì*.

Bộ dữ liệu không cần lớn tới tera byte. Trong trường hợp sumo, toàn bộ bộ dữ liệu chứa đựng ít bit hơn so với một bức ảnh kỹ thuật số điển hình ngày nay. Nhưng vì phân tích dữ-liệu-lớn, nó xem xét nhiều hơn so với một mẫu ngẫu nhiên điển hình. Khi nói về dữ liệu lớn, chúng ta có ý nói "lớn" trong tương đối hơn là trong tuyệt đối: tương đối so với tập hợp toàn diện của dữ liệu.

Trong một thời gian dài, lấy mẫu ngẫu nhiên là một cách đi tắt hiệu quả. Nó làm cho việc phân tích các bài toán dữ liệu lớn nhất thành khả hiện trong thời kỳ tiền kỹ thuật số. Nhưng cũng giống như khi chuyển đổi một tấm ảnh hoặc bài hát kỹ thuật số vào một tập tin nhỏ hơn, thông tin bị mất khi lấy mẫu. Việc có đầy đủ (hoặc gần đầy đủ) tập dữ liệu sẽ tạo điều kiện tốt hơn để khám phá, để nhìn vào dữ liệu từ các góc độ khác nhau hoặc để xem xét kỹ hơn các khía cạnh nhất định của nó. Một cách so sánh phù hợp có thể là máy ảnh Lytro, không chỉ chụp một mặt phẳng ánh sáng đơn nhất, như với những máy ảnh thông thường, mà chụp tất cả các tia từ toàn bộ trường ánh sáng, khoảng 11 triệu phần tử. Người chụp hình sau đó có thể quyết định tập trung vào yếu tố nào của ảnh trong tập tin kỹ thuật số. Như vậy, không cần phải tập trung ngay từ đầu, bởi việc thu thập tất cả các thông tin cho phép có thể làm điều đó về sau.



Phim minh họa máy ảnh Lytro

Tương tự như vậy, vì dữ liệu lớn dựa trên tất cả các thông tin, hoặc nhiều thông tin nhất có thể, nên nó cho phép chúng ta nhìn vào các chi tiết hoặc thử nghiệm các phân tích mới mà không ngại rủi ro bị mất chất lượng. Chúng ta có thể kiểm tra các giả thuyết mới ở nhiều cấp độ chi tiết. Tính chất này chính là thứ cho phép chúng ta thấy được sự gian lận trong các trận đấu vật sumo, theo dõi sự lây lan của virus cúm theo vùng, và chống ung thư bằng cách nhắm vào một phần chính xác trên DNA của bệnh nhân. Nó cho phép chúng ta làm việc ở một mức độ rõ ràng tuyệt vời.

Tất nhiên, việc sử dụng tất cả các dữ liệu thay vì một mẫu không phải là luôn luôn cần thiết. Chúng ta vẫn sống trong một thế giới có nguồn lực hạn chế. Nhưng trong ngày càng nhiều trường hợp

thì việc sử dụng tất cả các dữ liệu có trong tay tỏ ra hợp lý, và làm như vậy là khả thi trong khi trước đây thì không.

Một trong các lĩnh vực chịu tác động mạnh nhất bởi N = tất cả là khoa học xã hội. Chúng đã mất đi độc quyền trong việc làm nên ý nghĩa cho dữ liệu thực nghiệm xã hội, khi phân tích dữ liệu lớn thay thế các chuyên gia khảo sát có tay nghề cao trong quá khứ. Các ngành khoa học xã hội chủ yếu dựa trên các nghiên cứu lấy mẫu và bảng câu hỏi. Nhưng khi dữ liệu được thu thập một cách thụ động trong khi mọi người tiếp tục làm những gì họ vẫn thường làm, thì những định kiến cũ liên quan đến lấy mẫu và bảng câu hỏi biến mất. Bây giờ chúng ta có thể thu thập được những thông tin mà ta không thể thu thập nổi trước đây, đó có thể là những mối quan hệ tiết lộ qua các cuộc gọi điện thoại di động hay những cảm xúc bộc lộ qua tweet. Quan trọng hơn, sự cần thiết phải lấy mẫu biến mất.

Albert-László Barabási, một trong những chuyên gia uy tín hàng đầu thế giới về khoa học lý thuyết mạng, muốn nghiên cứu sự tương tác giữa con người ở quy mô của toàn bộ dân số. Vì vậy, ông và các đồng nghiệp đã khảo sát các bản lưu ẩn danh của các cuộc gọi điện thoại di động từ một nhà điều hành phục vụ khoảng một phần năm dân số của một quốc gia châu Âu không xác định - tất cả các bản lưu trong thời gian bốn tháng. Đó là phân tích mạng lưới đầu tiên ở mức độ toàn xã hội, sử dụng một bộ dữ liệu trong tinh thần của N = tất cả. Một quy mô lớn như vậy - xem xét tất cả các cuộc gọi giữa hàng triệu người - đã tạo ra những hiểu biết mới không thể phát hiện được bằng bất kỳ phương cách nào khác.

Điều thú vị là trái ngược với các nghiên cứu nhỏ hơn, nhóm nghiên cứu phát hiện ra rằng nếu loại bỏ khỏi mạng lưới những người có nhiều liên kết ngay trong cộng đồng này, thì mạng xã hội còn lại sẽ giảm chất lượng nhưng không sụp đổ. Ngược lại, khi những người có liên kết bên ngoài cộng đồng trực tiếp này

được mang ra khỏi mạng, thì mạng xã hội đột ngột tan rã, giống như cấu trúc của nó bị khóa. Đó là một kết quả quan trọng, nhưng phần nào bất ngờ. Ai có thể nghĩ rằng những người có rất nhiều bạn bè thân thiết lại ít quan trọng hơn nhiều đối với sự ổn định của cấu trúc mạng so với những người có quan hệ với những người ở xa hơn? Nó cho thấy rằng sự đa dạng trong một nhóm và trong xã hội nói chung có một tầm quan trọng đặc biệt.

Chúng ta có xu hướng nghĩ về mẫu thống kê như một loại nền tảng bất biến, giống như các nguyên lý của hình học, hay các định luật của lực hấp dẫn. Tuy nhiên, khái niệm này mới ra đời chưa đầy một thế kỷ, và nó được phát triển để giải quyết một bài toán đặc biệt tại một thời điểm đặc biệt dưới những hạn chế cụ thể về công nghệ. Những hạn chế này không còn tồn tại với cùng mức độ nữa. Việc cố đạt được một mẫu ngẫu nhiên trong thời đại của dữ liệu lớn cũng giống như việc níu chặt một cây roi ngựa trong thời đại của xe hơi. Chúng ta vẫn có thể áp dụng cách lấy mẫu trong những hoàn cảnh nhất định, nhưng nó không cần, và sẽ không là cách chiếm ưu thế để chúng ta phân tích các bộ dữ liệu lớn. Càng ngày tất cả chúng ta sẽ càng nhắm đến điều đó.

3. HỖN ĐỘN

NGÀY CÀNG CÓ NHIỀU BỐI CẢNH, trong đó việc sử dụng tất cả các dữ liệu có sẵn là khả thi. Tuy nhiên nó đi kèm với chi phí. Tăng khối lượng sẽ mở cánh cửa cho sự thiếu chính xác. Điều chắc chắn là những số liệu sai sót và bị hỏng đã luôn luôn len lỏi vào các bộ dữ liệu. Chúng ta đã luôn luôn xem chúng như những rắc rối và cố gắng loại bỏ chúng, một phần vì chúng ta có thể làm được như vậy. Những gì chúng ta chưa bao giờ muốn làm là xem chúng như điều không thể tránh khỏi và học cách sống chung với chúng. Đây là một trong những thay đổi cơ bản khi chuyển từ dữ liệu nhỏ sang dữ liệu lớn.

Trong thế giới của dữ liệu nhỏ, giảm sai sót và đảm bảo chất lượng cao của dữ liệu là một động lực tự nhiên và cần thiết. Vì chỉ thu thập được một ít thông tin, chúng ta phải bảo đảm rằng những con số đã được cố gắng ghi lại là chính xác nhất có thể. Nhiều thế hệ các nhà khoa học đã tối ưu hóa các công cụ để các phép đo đạc của họ ngày càng chính xác hơn, dù là để xác định vị trí của các thiên thể hay kích thước của các đối tượng dưới kính hiển vi. Trong thế giới lấy mẫu, nỗi ám ảnh với sự chính xác thậm chí còn nặng nề hơn. Việc phân tích chỉ một số lượng hạn chế các điểm dữ liệu có nghĩa là lỗi có thể được khuếch đại, có khả năng làm giảm tính chính xác của kết quả tổng thể.

Trong phần lớn lịch sử, những thành quả cao nhất của loài người xuất hiện từ việc chinh phục thế giới bằng cách đo lường nó. Việc tìm kiếm sự chính xác bắt đầu tại châu Âu vào giữa thế kỷ thứ mười ba, khi các nhà thiên văn học và các học giả đã gánh vác việc định lượng thời gian và không gian một cách chính xác hơn bao giờ hết - đó là “đo lường hiện thực”, theo như lời của nhà sử học Alfred Crosby.

Nếu có thể đo lường một hiện tượng thì người ta tin rằng có thể hiểu được nó. Sau này, đo lường đã được gắn liền với phương

pháp quan sát và giải thích khoa học: khả năng định lượng, ghi nhận, và trình bày các kết quả có thể tái lập được. "Đo lường là để hiểu biết", Lord Kelvin đã phát biểu như vậy. Nó đã trở thành một cơ sở của quyền lực. "Hiểu biết là quyền lực", Francis Bacon nhận định. Đồng thời, các nhà toán học và những người sau này được gọi là kế toán đã phát triển những phương pháp để có thể thực hiện việc thu thập, ghi nhận, và quản lý dữ liệu một cách chính xác.

Đến thế kỷ XIX, Pháp - lúc đó là quốc gia hàng đầu thế giới về khoa học - đã phát triển một hệ thống các đơn vị đo lường được xác định chính xác để nắm bắt không gian, thời gian, và nhiều thứ khác nữa, và bắt đầu đề nghị các quốc gia khác cũng áp dụng cùng một tiêu chuẩn. Thậm chí họ đã đưa ra những đơn vị mẫu được quốc tế công nhận dùng để đo lường trong các hiệp ước quốc tế. Đó là đỉnh điểm của thời đại về đo lường. Chỉ một nửa thế kỷ sau đó, vào những năm 1920, các khám phá của cơ học lượng tử đã làm tan vỡ mãi mãi ước mơ của đo lường toàn diện và hoàn hảo. Tuy nhiên, bên ngoài phạm vi tương đối nhỏ của các nhà vật lý, thì suy nghĩ hướng tới đo lường một cách hoàn hảo vẫn tiếp tục đối với các kỹ sư và các nhà khoa học. Trong thế giới kinh doanh nó thậm chí còn được mở rộng, khi các ngành toán học và thống kê bắt đầu gây ảnh hưởng đến tất cả các lĩnh vực thương mại.

Tuy nhiên, trong nhiều tình huống mới nảy sinh ngày hôm nay, việc cho phép sự không chính xác - sự hỗn độn - có thể là một tính năng tích cực, chứ không phải là một thiếu sót. Nó là một sự cân bằng. Để bù đắp cho sự nới lỏng về tiêu chuẩn với các lỗi cho phép, người ta có thể có được nhiều dữ liệu hơn. Nó không chỉ mang ý nghĩa *"nhiều hơn* thì *tốt hơn",* mà thật ra đôi khi nó sẽ là *"nhiều hơn* thì tốt hơn cả *tốt hơn".*

Chúng ta phải đối mặt với nhiều loại hỗn độn khác nhau. Hỗn độn có thể mang một ý nghĩa đơn giản là khả năng sai sót tăng

lên khi bạn thêm điểm dữ liệu. Khi số lượng tăng lên gấp hàng ngàn lần thì khả năng một số trong đó có thể sai cũng tăng lên. Nhưng bạn cũng có thể làm tăng hỗn độn bằng cách kết hợp nhiều loại thông tin khác nhau từ các nguồn khác nhau, không luôn luôn tương thích với nhau một cách hoàn hảo. Ví dụ, nếu sử dụng phần mềm nhận dạng giọng nói để mô tả các khiếu nại đến một trung tâm tiếp nhận cuộc gọi, và so sánh dữ liệu này với khi dùng nhân viên để xử lý các cuộc gọi, người ta có thể có được một sự hình dung thực tế, tuy không hoàn hảo nhưng hữu ích. Hỗn độn cũng có thể tham chiếu tới sự không thống nhất định dạng, trong đó các dữ liệu cần được "làm sạch" trước khi được xử lý. Ví dụ chuyên gia dữ liệu lớn DJ Patil nhận xét từ viết tắt IBM có rất nhiều cách diễn đạt, như hoặc Phòng thí nghiệm T.J. Watson, hoặc International Business Machines. Và hỗn độn có thể phát sinh khi chúng ta trích xuất hoặc xử lý dữ liệu, vì khi làm như vậy, chúng ta đang chuyển đổi nó, biến nó thành một cái gì đó khác, chẳng hạn như khi chúng ta thực hiện phân tích cảm nghĩ các tin nhắn Twitter để dự đoán doanh thu phòng vé của Hollywood. Chính bản thân sự hỗn độn cũng mang tính hỗn độn.

Giả sử chúng ta cần đo nhiệt độ trong một vườn nho. Nếu chúng ta chỉ có một cảm biến nhiệt độ cho toàn bộ lô đất, chúng ta phải chắc chắn rằng nó chính xác và hoạt động được tại mọi thời điểm: sự hỗn độn không được tồn tại. Ngược lại, nếu chúng ta có một cảm biến cho mỗi cây trong vườn hàng trăm cây nho, chúng ta có thể sử dụng những cảm biến rẻ hơn, ít phức tạp hơn (miễn là chúng không phát sinh một sai số có hệ thống). Rất có thể là tại một số thời điểm, một vài cảm biến sẽ báo dữ liệu không chính xác, tạo ra một bộ dữ liệu ít chính xác, hoặc "hỗn độn" hơn so với bộ dữ liệu từ một cảm biến chính xác đơn nhất. Bất kỳ phép đọc cụ thể nào đó cũng đều có thể không chính xác, nhưng tổng hợp của nhiều phép đọc sẽ cung cấp một bức tranh toàn diện hơn. Bởi vì bộ dữ liệu này bao gồm nhiều điểm dữ liệu hơn, nó cung cấp giá trị lớn hơn nhiều và có thể bù đắp cho sự hỗn

động của nó.

Bây giờ giả sử chúng ta tăng tần số các lần đọc cảm biến. Nếu đo mỗi phút một lần, chúng ta có thể khá chắc chắn rằng trình tự mà các dữ liệu đến sẽ hoàn toàn theo thứ tự thời gian. Nhưng nếu chúng ta thay đổi, đọc đến mười hay một trăm lần trong một giây, thì độ chính xác của trình tự có thể trở nên không chắc chắn. Khi thông tin đi qua mạng, một bản ghi có thể bị trì hoãn và đến lệch trình tự, hoặc đơn giản là có thể bị mất. Thông tin sẽ ít chính xác đi một chút, nhưng khối lượng lớn sẽ khiến cho khả năng từ bỏ sự chính xác nghiêm ngặt trở nên thích đáng.

Trong ví dụ đầu tiên, chúng ta đã hy sinh tính chính xác của mỗi điểm dữ liệu cho chiều rộng, và ngược lại chúng ta nhận được tính chi tiết mà bình thường chúng ta có thể đã không nhìn thấy. Trong trường hợp thứ hai, chúng ta đã từ bỏ sự chính xác cho tần số, và ngược lại, chúng ta thấy sự thay đổi mà bình thường chúng ta đã phải bỏ qua. Mặc dù có thể khắc phục những sai sót nếu chúng ta đầu tư đủ nguồn lực vào đó - xét cho cùng, mỗi giây có tới 30.000 giao dịch xảy ra trên Thị trường Chứng khoán New York, nơi trình tự chính xác là vấn đề rất quan trọng - trong nhiều trường hợp, việc chấp nhận lỗi thay vì cố gắng ngăn chặn nó lại tỏ ra hiệu quả hơn.

Ví dụ, chúng ta có thể chấp nhận sự hỗn độn để đổi lấy quy mô. Như Forrester, một nhà tư vấn công nghệ, đã nói: “Đôi khi hai cộng với hai có thể bằng 3,9, và như vậy là đủ tốt”. Tất nhiên dữ liệu không được phép sai hoàn toàn, nhưng chúng ta sẵn sàng hy sinh một chút trong sự chính xác để đổi lại hiểu biết về xu hướng chung. Dữ liệu lớn biến đổi các con số thành một cái gì đó mang tính xác suất nhiều hơn là tính chính xác. Thay đổi này sẽ cần rất nhiều để làm quen, và nó cũng đi kèm với những vấn đề riêng của nó, mà chúng ta sẽ xem xét sau trong cuốn sách. Nhưng bây giờ, hãy đơn giản lưu ý rằng chúng ta thường sẽ cần đón nhận lấy sự hỗn độn khi chúng ta tăng quy mô.

Người ta thấy một sự thay đổi tương tự về tầm quan trọng của việc có nhiều dữ liệu hơn, liên quan tới những cải tiến khác trong điện toán. Mọi người đều biết sức mạnh xử lý đã tăng lên ra sao trong những năm qua như dự đoán của Định luật Moore, phát biểu rằng số lượng bán dẫn trên một chip tăng gấp đôi khoảng mỗi hai năm. Sự cải tiến liên tục này đã làm máy tính nhanh hơn và bộ nhớ phong phú hơn. Ít người trong chúng ta biết rằng hiệu suất của các thuật toán điều khiển nhiều hệ thống của chúng ta cũng đã tăng lên - trong nhiều lĩnh vực, với mức tăng còn hơn cả mức cải thiện của các bộ xử lý theo Định luật Moore. Tuy nhiên, nhiều lợi ích cho xã hội từ dữ liệu lớn lại xảy ra không phải vì các chip nhanh hơn hay vì các thuật toán tốt hơn, mà vì có nhiều dữ liệu hơn.

Ví dụ, các thuật toán chơi cờ chỉ thay đổi chút ít trong vài thập kỷ qua, bởi các quy tắc của cờ vua đã được biết đầy đủ và bị giới hạn một cách chặt chẽ. Lý do các chương trình cờ vua ngày nay chơi tốt hơn trước đây rất nhiều là một phần bởi chúng chơi cờ tàn tốt hơn. Và chúng làm được điều đó đơn giản chỉ vì các hệ thống được cung cấp nhiều dữ liệu hơn. Thực tế, cờ tàn với sáu hoặc ít quân hơn còn lại trên bàn cờ đã được phân tích một cách hoàn toàn đầy đủ và tất cả các bước đi có thể (N = tất cả) đã được thể hiện trong một bảng lớn, khi không nén sẽ lấp đầy hơn một tera byte dữ liệu. Điều này cho phép các máy tính có thể chơi cờ tàn một cách hoàn hảo. Không bao giờ con người có thể chơi thắng được hệ thống.

Ý nghĩa của lập luận rằng “có nhiều dữ liệu hơn sẽ hiệu quả hơn việc có các thuật toán tốt hơn” đã được thể hiện mạnh mẽ trong lĩnh vực xử lý ngôn ngữ tự nhiên: cách các máy tính học phân tích cú pháp các từ như chúng ta sử dụng chúng trong giao tiếp hàng ngày. Khoảng năm 2000, các nhà nghiên cứu Michele Banko và Eric Brill của Microsoft tìm kiếm một phương pháp để cải thiện bộ kiểm tra ngữ pháp, một thành phần của chương trình Microsoft Word. Họ không chắc liệu sẽ hữu ích hơn nếu

dành nỗ lực của mình vào việc cải thiện các thuật toán sẵn có, hay tìm kiếm các kỹ thuật mới, hay bổ sung thêm những tính năng phức tạp hơn. Trước khi đi theo bất kỳ con đường nào, họ quyết định xem xét những gì sẽ xảy ra khi họ cung cấp thêm rất nhiều dữ liệu cho các phương pháp hiện có. Hầu hết các thuật toán học tập của máy dựa trên những tập sao lục văn bản đạt tới một triệu từ hoặc ít hơn. Banko và Brill lấy bốn thuật toán thông thường và cung cấp nhiều dữ liệu hơn ở ba cấp độ khác nhau: 10 triệu từ, sau đó 100 triệu, và cuối cùng là 1 tỷ từ.

Kết quả thật đáng kinh ngạc. Khi có nhiều dữ liệu đi vào, hiệu suất của tất cả bốn loại thuật toán đều được cải thiện một cách đáng kể. Trong thực tế, một thuật toán đơn giản hoạt động kém hiệu quả nhất với một nửa triệu từ lại hoạt động tốt hơn những thuật toán khác khi có một tỷ từ. Độ chính xác của nó đã tăng từ 75 phần trăm lên trên 95 phần trăm. Ngược lại, thuật toán làm việc tốt nhất với ít dữ liệu lại hoạt động kém nhất với lượng dữ liệu lớn hơn, mặc dù cũng giống như những thuật toán khác nó được cải thiện rất nhiều, tăng từ khoảng 86 phần trăm lên 94 phần trăm chính xác. "Những kết quả này cho thấy chúng ta có thể nên xem xét lại sự cân bằng giữa việc tiêu tốn thời gian và tiền bạc vào phát triển thuật toán so với việc chi tiêu vào phát triển ngữ liệu", Banko và Brill đã viết trong một tài liệu nghiên cứu của họ về chủ đề này.

Vậy là *nhiều hơn* đã thắng *ít hơn*. Và đôi khi *nhiều hơn* còn thắng cả *thông minh hơn*. Còn sự hỗn độn thì sao? Một vài năm sau khi Banko và Brill đào bới tất cả những dữ liệu này, các nhà nghiên cứu đối thủ Google đã suy nghĩ dọc theo dòng tương tự - nhưng với quy mô lớn hơn. Thay vì thử các thuật toán với một tỷ từ, họ đã sử dụng một ngàn tỷ từ. Google làm điều này không phải để phát triển một bộ kiểm tra ngữ pháp, nhưng để giải quyết một trở ngại thậm chí còn phức tạp hơn: dịch thuật. Cái gọi là dịch máy đã ở trong tầm nhìn của những nhà tiên phong máy tính ngay từ buổi bình minh của tính toán trong những năm 1940,

khi các thiết bị được làm bằng đèn chân không và chứa đầy cả một căn phòng. Ý tưởng được nâng lên thành cấp bách đặc biệt trong Chiến tranh Lạnh, khi Hoa Kỳ thu được một lượng lớn tư liệu viết và nói tiếng Nga nhưng thiếu nhân lực để dịch nó một cách nhanh chóng.

Lúc đầu, các nhà khoa học máy tính đã lựa chọn một sự kết hợp của các quy tắc ngữ pháp và một từ điển song ngữ. Một máy tính IBM đã dịch sáu mươi câu từ tiếng Nga sang tiếng Anh vào năm 1954, sử dụng 250 cặp từ trong từ vựng của máy tính và sáu quy tắc ngữ pháp. Kết quả rất hứa hẹn. *“Mi pyeryedayem mislyỉ posryedstvom ryechyi”,* được nhập vào máy IBM 701 qua bìa đục lỗ, và đầu ra có “Chúng tôi truyền suy nghĩ bằng lời nói”. Sáu mươi câu đã được “dịch trơn tru”, theo một thông cáo báo chí của IBM kỷ niệm sự kiện này. Giám đốc chương trình nghiên cứu, Leon Dostert của Đại học Georgetown, dự đoán rằng dịch máy sẽ trở thành “thực tế” trong “năm, hay có thể là ba năm nữa”. Nhưng thành công ban đầu hóa ra lại tạo một sự hiểu lầm khá sâu sắc. Đến năm 1966 một ủy ban của các đại thụ trong làng dịch máy đã phải thừa nhận thất bại. Vấn đề khó hơn họ tưởng. Dạy máy tính dịch là dạy chúng không chỉ các quy tắc, mà cả các trường hợp ngoại lệ nữa. Dịch không chỉ là ghi nhớ và nhớ lại, nó là về việc chọn những từ thích hợp từ nhiều lựa chọn thay thế. Liệu “bonjour” có thực sự là “chào buổi sáng”? Hay đó là “ngày tốt”, hay “xin chào”, hay “hi”? Câu trả lời là “còn tùy”.

Cuối những năm 1980, các nhà nghiên cứu tại IBM đã có một ý tưởng mới lạ. Thay vì cố gắng nạp những quy tắc ngôn ngữ rõ ràng vào máy tính cùng với một từ điển, họ đã quyết định để cho máy tính sử dụng xác suất thống kê để tính toán xem từ hoặc câu nào trong một ngôn ngữ là thích hợp nhất với từ hoặc câu trong một ngôn ngữ khác. Trong những năm 1990 dự án Candide của IBM đã sử dụng các văn bản quốc hội Canada công bố bằng tiếng Pháp và tiếng Anh trong vòng mười năm - khoảng ba triệu cặp câu. Do chúng là văn bản chính thức, nên các bản

dịch đã được thực hiện với chất lượng đặc biệt cao. Và theo các tiêu chuẩn lúc đó, số lượng dữ liệu là rất lớn. Dịch máy thống kê, như kỹ thuật này được biết đến, đã khéo léo biến những thách thức của dịch thuật thành một bài toán lớn của toán học. Và nó dường như thành công. Đột nhiên, dịch máy trở thành tốt hơn rất nhiều. Tuy nhiên, sau thành công của bước nhảy vọt về khái niệm, IBM chỉ thu được những cải thiện nhỏ mặc dù phải ném ra rất nhiều tiền. Cuối cùng IBM đã dừng dự án.

Nhưng chưa đầy một thập kỷ sau đó, vào năm 2006, Google đã nhảy vào dịch thuật, như một phần của nhiệm vụ "tổ chức thông tin của thế giới và làm cho chúng trở thành có thể tiếp cận được và hữu ích một cách phổ dụng". Thay vì dịch các trang văn bản thành hai ngôn ngữ, Google tự giúp mình với một bộ dữ liệu lớn hơn nhưng cũng hỗn độn hơn nhiều: toàn bộ mạng Internet toàn cầu và nhiều hơn nữa. Hệ thống của Google đã thu lượm bất kể bản dịch nào có thể tìm thấy, để huấn luyện máy tính. Chúng bao gồm các trang web của các công ty viết ở nhiều ngôn ngữ khác nhau, các bản dịch đồng nhất của các văn bản chính thức, và các báo cáo của các tổ chức liên chính phủ như Liên hợp quốc và Liên minh châu Âu. Thậm chí các bản dịch sách từ dự án sách của Google cũng được thu nhận. Trong khi Candide sử dụng ba triệu câu được dịch một cách cẩn thận, thì hệ thống của Google khai thác hàng tỷ trang các bản dịch rất khác nhau về chất lượng, theo người đứng đầu của Google Translate, Franz Josef Och, một trong những chuyên gia uy tín nhất trong lĩnh vực này. Hàng nghìn tỷ từ đã được chuyển thành 95 tỷ câu tiếng Anh, mặc dù chất lượng không rõ ràng.

Bất chấp sự hỗn độn của đầu vào, dịch vụ của Google hoạt động tốt nhất. Các bản dịch của nó là chính xác hơn so với của các hệ thống khác (mặc dù vẫn còn kém). Và nó phong phú hơn rất nhiều. Vào giữa năm 2012 bộ dữ liệu của nó bao gồm hơn 60 ngôn ngữ. Nó thậm chí có thể chấp nhận nhập văn bản vào bằng giọng nói trong 14 ngôn ngữ để dịch. Và vì nó xử lý ngôn ngữ

đơn giản như là dữ liệu hỗn độn để đánh giá xác suất, nó thậm chí có thể dịch giữa các ngôn ngữ, chẳng hạn như giữa tiếng Hindi và Catalan, mà trong đó có rất ít bản dịch trực tiếp để phát triển hệ thống. Trong những trường hợp này, nó sử dụng tiếng Anh như một cầu nối. Và nó linh hoạt hơn nhiều so với những cách tiếp cận khác, vì nó có thể thêm và bớt các từ qua kinh nghiệm chúng được hay không được sử dụng.

Lý do hệ thống dịch thuật của Google hoạt động tốt không phải vì nó có một thuật toán thông minh hơn. Nó hoạt động tốt bởi vì tác giả của nó, như Banko và Brill tại Microsoft, nạp vào nhiều dữ liệu hơn - và không chỉ dữ liệu chất lượng cao. Google đã có thể sử dụng một bộ dữ liệu *hàng chục ngàn* lần lớn hơn hơn Candide của IBM vì nó chấp nhận sự hỗn độn. Cả nghìn tỷ ngữ liệu Google phát hành năm 2006 được biên soạn từ đủ thứ, kể cả đồ tạp nham và đồ bỏ đi của Internet - có thể nói là "dữ liệu thượng vàng hạ cám". Đây là các "tập huấn luyện" để hệ thống có thể tính toán xác suất, ví dụ một từ trong tiếng Anh đi tiếp sau một từ khác. Đó là một mong ước xa vời của ông tổ trong lĩnh vực này, dự án Brown Corpus nổi tiếng vào những năm 1960, đã tập hợp được tổng cộng một triệu từ tiếng Anh. Việc sử dụng bộ dữ liệu lớn hơn cho phép những bước tiến lớn trong xử lý ngôn ngữ tự nhiên, mà các hệ thống nhận dạng tiếng nói và dịch máy dựa vào. "Mô hình đơn giản và rất nhiều dữ liệu thắng thế những mô hình phức tạp hơn nhưng dựa trên ít dữ liệu hơn", chuyên gia trí tuệ nhân tạo của Google, Peter Norvig và các đồng nghiệp đã viết như vậy trong một bài báo có tựa đề "Hiệu quả phi lý của dữ liệu" ("The Unreasonable effectiveness of Data"): "Có thể nói ngữ liệu này là một bước lùi từ Brown Corpus: nó được lấy từ các trang web chưa được hiệu đính và do đó chứa những câu chưa đầy đủ, lỗi chính tả, lỗi ngữ pháp, và tất cả các loại lỗi khác. Nó không được chú thích cẩn thận với những thẻ bài được chỉnh sửa. Nhưng việc nó lớn hơn một triệu lần so với Brown Corpus đã đủ bù đắp cho những hạn chế này".



Phim minh họa hệ thống GoogleTranslate

Nhiều hơn thắng thế tốt hơn

Hỗn độn rất khó được các nhà phân tích mẫu thông thường chấp nhận, vì họ là những người cả đời đã tập trung vào việc ngăn chặn và xóa bỏ sự hỗn độn. Họ làm việc chăm chỉ để giảm tỷ lệ lỗi khi thu thập mẫu, và để kiểm tra các mẫu nhằm loại bỏ các thành kiến tiềm ẩn trước khi công bố kết quả của mình. Họ sử dụng nhiều chiến lược giảm lỗi, trong đó có việc đảm bảo mẫu được thu thập theo một giao thức chính xác và bởi các chuyên gia được huấn luyện đặc biệt. Những chiến lược như vậy rất tốn kém khi thực hiện, ngay cả đối với số lượng hạn chế các điểm dữ liệu, và chúng hầu như không khả thi cho dữ liệu lớn. Không chỉ vì nó quá đắt, mà còn vì những tiêu chuẩn chính xác của việc tập hợp là khó có thể đạt được một cách nhất quán ở quy mô như

vậy. Thậm chí loại bỏ sự tương tác của con người cũng sẽ không giải quyết được vấn đề.

Di chuyển vào một thế giới của dữ liệu lớn sẽ đòi hỏi chúng ta thay đổi tư duy về giá trị của sự chính xác. Việc áp dụng tư duy thông thường của đo lường vào thế giới kỹ thuật số được kết nối của thế kỷ XXI đồng nghĩa với bỏ lỡ một điểm quan trọng. Như đã đề cập trước đây, nỗi ám ảnh với tính chính xác là một tạo tác của thời đại analog. Khi dữ liệu thưa thớt, mỗi điểm dữ liệu đều quan trọng, và do đó người ta thận trọng tránh để bất kỳ điểm dữ liệu nào gây sai lệch cho việc phân tích. Ngày nay chúng ta không còn sống trong tình trạng bị đói thông tin. Trong khi làm việc với các bộ dữ liệu ngày càng toàn diện hơn, không chỉ thâu tóm một mảnh nhỏ của hiện tượng mà nhiều hơn hoặc tất cả, chúng ta không cần lo lắng quá nhiều về việc các điểm dữ liệu riêng lẻ gây ra sai lệch cho phân tích tổng thể. Thay vì nhắm tới sự chính xác từng tí một với chi phí ngày càng cao, chúng ta đang tính toán với sự hỗn độn trong tâm thức.

Hãy xem các cảm biến đã thâm nhập vào nhà máy như thế nào. Tại nhà máy lọc dầu Cherry Point ở Blaine, bang Washington, các bộ cảm biến không dây được cài đặt khắp nơi, tạo thành một lưới vô hình thu thập những lượng lớn dữ liệu trong thời gian thực. Môi trường nhiệt độ cao và máy móc điện tử có thể làm sai lệch các phép đọc, dẫn tới dữ liệu lộn xộn. Nhưng lượng thông tin khổng lồ được tạo ra từ các cảm biến, cả có dây và không dây, sẽ dung hòa cho những trục trặc này. Chỉ cần tăng tần số và số địa điểm đọc cảm biến là có thể thu được lợi thế lớn. Bằng cách đo sức căng trên đường ống ở tất cả các thời điểm chứ không phải chỉ tại những khoảng thời gian nhất định, BP biết được một số loại dầu thô ăn mòn nhiều hơn những loại khác - điều nó không thể phát hiện, và do đó không thể chống lại, khi bộ dữ liệu nhỏ hơn.

Khi số lượng dữ liệu lớn hơn nhiều và là một loại mới, độ chính

xác trong một số trường hợp không còn là mục tiêu, miễn là chúng ta có thể thấy được xu hướng chung. Việc chuyển sang một quy mô lớn làm thay đổi không chỉ sự mong đợi về độ chính xác mà cả khả năng thực tế để đạt được sự chính xác. Dù nó có vẻ phản lại trực giác lúc đầu, việc xử lý dữ liệu như một cái gì đó không hoàn hảo và không chính xác cho phép chúng ta đưa ra dự báo tốt hơn, và do đó hiểu biết thế giới của chúng ta tốt hơn.

Nên lưu ý rằng hỗn độn không phải là đặc tính vốn có của dữ liệu lớn. Thay vào đó, nó là một chức năng của sự không hoàn hảo của các công cụ chúng ta sử dụng để đo lường, ghi nhận và phân tích thông tin. Nếu công nghệ bằng cách nào đó trở nên hoàn hảo, thì vấn đề của sự không chính xác sẽ biến mất. Nhưng một khi nó còn là không hoàn hảo, thì sự hỗn độn là một thực tế mà chúng ta phải đối mặt. Và nhiều khả năng nó sẽ còn tồn tại với chúng ta trong một thời gian dài. Nỗ lực để tăng độ chính xác thường sẽ không có ý nghĩa kinh tế, bởi giá trị của việc có những lượng dữ liệu lớn hơn sẽ hấp dẫn hơn. Giống như các nhà thống kê trong kỷ nguyên trước đây đã gạt sang một bên mối quan tâm của họ tới những kích thước mẫu lớn hơn, để ủng hộ sự ngẫu nhiên hơn, chúng ta có thể sống với một chút không chính xác để đổi lấy nhiều dữ liệu hơn.

Dự án Billion Prices cung cấp một trường hợp khá hấp dẫn. Mỗi tháng Cục Thống kê Lao động Mỹ công bố chỉ số giá tiêu dùng, hay CPI, được sử dụng để tính toán tỷ lệ lạm phát. Chỉ số liệu này là rất quan trọng cho các nhà đầu tư và doanh nghiệp. Cục Dự trữ Liên bang xem xét nó khi quyết định nên tăng hoặc giảm lãi suất. Lương cơ bản của các công ty tăng khi có lạm phát. Chính phủ liên bang sử dụng nó để điều chỉnh khoản thanh toán như trợ cấp an sinh xã hội và lãi suất trả cho những trái phiếu nhất định.

Để có được chỉ số này, Cục Thống kê Lao động sử dụng hàng trăm nhân viên để gọi điện, gửi fax, ghé thăm các cửa hàng và

văn phòng tại 90 thành phố trên toàn quốc và báo cáo lại khoảng 80.000 mức giá về tất cả mọi thứ từ giá cà chua tới giá đi taxi. Để có nó, người ta phải chi ra khoảng 250 triệu USD một năm. Với số tiền này, dữ liệu được gọn gàng, sạch sẽ và trật tự. Nhưng tại thời điểm các con số được công bố, chúng đã chậm mất vài tuần. Như cuộc khủng hoảng tài chính năm 2008 cho thấy, một vài tuần có thể là một sự chậm trễ khủng khiếp. Những người ra quyết định cần truy cập nhanh hơn đến các số liệu lạm phát để ứng phó với nó tốt hơn, nhưng họ không thể nhận được chúng với những phương pháp thông thường tập trung vào lấy mẫu và coi trọng sự chính xác.

Để đáp lại, hai nhà kinh tế tại Viện Công nghệ Massachusetts, Alberto Cavallo và Roberto Rigobon, đã tạo ra một phương pháp thay thế liên quan đến dữ-liệu-lớn, bằng cách đi theo một con đường hỗn độn hơn nhiều. Sử dụng phần mềm để thu thập dữ liệu web, họ đã có được nửa triệu giá của các sản phẩm được bán ở Mỹ mỗi ngày. Các thông tin là lộn xộn, và không phải tất cả các điểm dữ liệu thu thập được đều có thể dễ dàng so sánh với nhau. Nhưng bằng cách kết hợp bộ sưu tập dữ-liệu-lớn với phân tích thông minh, dự án đã có thể phát hiện một dao động giảm phát trong giá ngay sau khi ngân hàng Lehman Brothers đệ đơn xin phá sản vào tháng 9 năm 2008, trong khi những nơi phụ thuộc vào số liệu CPI chính thức đã phải chờ tới tháng Mười Một để nhìn thấy nó.

Dự án của MIT sau này đã tách ra thành một công ty thương mại gọi là PriceStats được các ngân hàng và những công ty khác sử dụng để đưa ra những quyết định kinh tế. Nó xử lý hàng triệu sản phẩm bán ra của hàng trăm nhà bán lẻ trong hơn 70 quốc gia mỗi ngày. Tất nhiên, các con số đòi hỏi phải có sự giải thích cẩn thận, nhưng chúng tốt hơn so với số liệu thống kê chính thức trong việc chỉ ra xu hướng lạm phát. Bởi vì có nhiều giá và các con số có sẵn trong thời gian thực, chúng cung cấp cho người ra quyết định một lợi thế đáng kể. (Phương pháp này cũng đóng vai

trò như một cách kiểm tra bên ngoài đáng tin cậy đối với các cơ quan thống kê quốc gia. Ví dụ, *The Economist* nghi ngờ phương pháp tính lạm phát của Argentina, vì vậy đã dùng các số liệu của PriceStats để thay thế.)

Áp dụng sự hỗn độn

Trong nhiều lĩnh vực công nghệ và xã hội, chúng ta đang nghiêng về ủng hộ sự nhiều hơn và sự hỗn độn chứ không phải sự ít hơn và sự chính xác. Hãy xem xét trường hợp của việc phân loại nội dung. Trong nhiều thế kỷ con người đã phát triển các nguyên tắc phân loại và chỉ số để lưu trữ và tìm kiếm tài liệu. Những hệ thống phân cấp này đã luôn luôn không hoàn hảo, như những ai từng quen thuộc với danh mục thẻ thư viện đều có thể đau đớn nhớ lại. Trong một thế giới dữ-liệu-nhỏ thì chúng hoạt động đủ tốt. Tuy nhiên khi tăng quy mô lên nhiều cấp độ, những hệ thống này, được cho là sắp xếp vị trí mọi thứ bên trong rất hoàn hảo, lại sụp đổ. Ví dụ, trong năm 2011 trang web chia sẻ hình ảnh Flickr có chứa hơn 6 tỷ hình ảnh từ hơn 75 triệu người sử dụng. Việc cố gắng gán nhãn cho từng bức ảnh theo những thể loại định trước đã tỏ ra vô ích. Liệu đã thực sự có một thể loại mang tên “Mèo trông giống như Hitler”?

Thay vào đó, nguyên tắc phân loại sạch được thay thế bằng cơ chế hỗn độn hơn nhưng linh hoạt hơn và dễ thích nghi hơn một cách xuất sắc với một thế giới luôn tiến hóa và thay đổi. Khi tải ảnh lên Flickr, chúng ta “gán thẻ (tag)” cho chúng. Có nghĩa là chúng ta gán một số bất kỳ các nhãn văn bản và sử dụng chúng để tổ chức và tìm kiếm các tư liệu. Thẻ được tạo ra và gán một cách đặc biệt: không có những danh mục tiêu chuẩn hóa, được định trước, không có phân loại sẵn để chúng ta phải tuân thủ. Thay vào đó, bất cứ ai cũng đều có thể thêm các thẻ mới bằng cách gõ chúng vào. Gắn thẻ đã nổi lên như tiêu chuẩn thực tế để phân loại nội dung trên Internet, được sử dụng trên các trang mạng xã hội như Twitter, các blog... Nó làm cho người sử dụng

dễ dàng di chuyển hơn trong sự bao la của nội dung các trang web - đặc biệt là cho những thứ như hình ảnh, phim, và âm nhạc không dựa trên văn bản nên việc tìm kiếm bằng từ không thể hoạt động được.

Tất nhiên, một số thẻ có thể bị viết sai chính tả, và những lỗi như vậy sẽ tạo ra sự không chính xác - không chỉ đối với chính dữ liệu, mà còn đối với việc chúng được tổ chức ra sao. Điều đó làm tổn thương tư duy truyền thống được rèn luyện trong sự chính xác. Nhưng bù lại cho sự hỗn độn trong cách chúng ta tổ chức các bộ sưu tập ảnh, chúng ta có được một vũ trụ phong phú hơn nhiều của các nhãn mác, và mở rộng ra, là một sự truy cập sâu hơn, rộng hơn tới các ảnh của chúng ta. Chúng ta có thể phối hợp các thẻ tìm kiếm để lọc các bức ảnh theo những cách không thể làm được trước đây. Sự thiếu chính xác vốn có trong gắn thẻ liên quan tới việc chấp nhận sự hỗn độn tự nhiên của thế giới. Nó là món thuốc giải độc cho các hệ thống chính xác hơn, vốn cố áp đặt tính tinh khiết sai lầm lên sự náo nhiệt của thực tế, giả vờ rằng tất cả mọi thứ dưới ánh mặt trời đều có thể được xếp ngay ngắn theo hàng và cột. Có nhiều thứ trên thiên đường và mặt đất hơn là những gì được mơ ước trong triết lý đó.

Nhiều trong số các trang web phổ biến nhất đã thể hiện rõ sự ưa thích tính thiếu chính xác hơn là sự kỳ vọng vào tính nghiêm cẩn. Khi người ta thấy một biểu tượng Twitter hay một nút "like" Facebook trên một trang web, nó cho thấy số lượng người đã nhấp chuột vào đó. Khi số lượng là nhỏ, mỗi cú nhấp chuột đều được hiển thị, như "63". Tuy nhiên, khi số lượng lớn lên, con số được hiển thị chỉ là một kiểu ước lượng, như "4K". Nó không có nghĩa là hệ thống không biết tổng số thực tế, mà chỉ vì khi quy mô tăng, thì việc cho thấy con số chính xác là ít quan trọng hơn. Bên cạnh đó, số lượng có thể thay đổi nhanh đến mức một con số cụ thể sẽ trở thành lạc hậu ngay vào thời điểm nó xuất hiện. Tương tự như vậy, Gmail của Google hiển thị thời gian của các tin nhắn mới nhất với độ chính xác cao, chẳng hạn như "11 phút

trước", nhưng với những thời lượng dài hơn thì nó tỏ ra thờ ơ, chẳng hạn như "2 giờ trước", cũng giống như Facebook và một số hệ thống khác.

Ngành công nghiệp tình báo kinh doanh và phần mềm phân tích từ lâu đã được xây dựng trên cơ sở hứa hẹn với khách hàng "một phiên bản duy nhất của sự thật" - lời đồn đại phổ biến của những năm 2000 từ các nhà cung cấp công nghệ trong lĩnh vực này. Các giám đốc điều hành đã sử dụng câu này không phải với sự mỉa mai. Và một số người vẫn còn làm như vậy. Bằng cách này, họ cho rằng tất cả những ai truy cập các hệ thống công nghệ thông tin của công ty đều có thể thâm nhập vào cùng một dữ liệu; như vậy nhóm tiếp thị và nhóm bán hàng không cần phải tranh cãi xem ai có số liệu chính xác về khách hàng hay doanh số trước khi cuộc họp thậm chí bắt đầu. Mối bận tâm của họ có thể trở nên hòa hợp hơn nếu các số liệu và sự kiện là nhất quán - kiểu tư duy này cứ tiếp diễn như vậy.

Nhưng ý tưởng về "một phiên bản duy nhất của sự thật" là một yếu tố dễ dàng trở mặt. Chúng ta đang bắt đầu nhận thấy một phiên bản duy nhất của sự thật chẳng những không thể tồn tại, mà việc theo đuổi nó là một sự điên rồ. Để gặt hái những lợi ích của việc khai thác dữ liệu với quy mô, chúng ta phải chấp nhận sự hỗn độn như một điều hiển nhiên, chứ không phải một cái gì đó chúng ta nên cố gắng loại bỏ.

Thậm chí chúng ta đang nhìn thấy những đặc tính của sự không chính xác xâm nhập vào một trong những lĩnh vực ít cởi mở nhất đối với nó: thiết kế cơ sở dữ liệu. Các hệ thống cơ sở dữ liệu truyền thống đòi hỏi dữ liệu phải có cấu trúc và tính chính xác rất cao. Dữ liệu không chỉ đơn giản được lưu trữ, chúng được chia thành "bản ghi" có chứa các trường. Mỗi trường lưu trữ thông tin với một kiểu và một độ dài nhất định. Ví dụ nếu một trường có độ dài bảy chữ số, khi đó số lượng 10 triệu hoặc lớn hơn sẽ không thể ghi lại được. Hoặc nếu muốn nhập cụm từ

“không xác định” vào một trường cho số điện thoại cũng không thể được. Cấu trúc của cơ sở dữ liệu phải được thay đổi để có thể chấp nhận những mục kiểu này. Chúng ta vẫn phải đánh vật với những hạn chế như vậy trên máy tính và điện thoại thông minh của mình, khi phần mềm không chấp nhận các dữ liệu chúng ta muốn nhập.

Các chỉ số truyền thống cũng được xác định trước, và như vậy hạn chế những gì người ta có thể tìm kiếm. Khi thêm một chỉ số mới thì phải tạo lập lại từ đầu, rất tốn thời gian. Những cơ sở dữ liệu thông thường, còn gọi là cơ sở dữ liệu quan hệ, được thiết kế cho một thế giới trong đó dữ liệu là thưa thót, và do đó có thể và sẽ được sửa chữa cẩn thận. Đó là một thế giới mà các câu hỏi người ta muốn trả lời bằng cách sử dụng dữ liệu phải rõ ràng ngay từ đầu, để cơ sở dữ liệu được thiết kế nhằm trả lời chúng - và chỉ có chúng - một cách hiệu quả.

Tuy nhiên, quan điểm này của lưu trữ và phân tích ngày càng mâu thuẫn với thực tế. Ngày nay chúng ta có những lượng lớn dữ liệu với các loại và chất lượng khác nhau. Hiếm khi nó phù hợp với những phân loại được xác định trước một cách quy củ. Và các câu hỏi chúng ta muốn hỏi thường chỉ xuất hiện khi chúng ta thu thập và làm việc với các dữ liệu mình có.

Những thực tế này đã dẫn đến những thiết kế cơ sở dữ liệu mới mẻ phá vỡ các nguyên tắc cũ - những nguyên tắc của bản ghi và các trường được thiết đặt trước, phản ánh những phân cấp được xác định một cách quy củ của thông tin. Ngôn ngữ phổ biến nhất để truy cập cơ sở dữ liệu từ lâu đã là SQL, hoặc “ngôn ngữ truy vấn có cấu trúc”. Cái tên gợi lên sự cứng nhắc của nó. Nhưng sự thay đổi lớn trong những năm gần đây là hướng tới một cái gì đó gọi là NoSQL, không đòi hỏi một cấu trúc bản ghi cài đặt sẵn để làm việc. Nó chấp nhận dữ liệu với kiểu và kích thước khác nhau và giúp tìm kiếm chúng thành công. Để đổi lại việc cho phép sự hỗn độn về cấu trúc, những thiết kế cơ sở dữ liệu này đòi hỏi

nhiều tài nguyên xử lý và dung lượng lưu trữ hơn. Tuy nhiên, đó là một sự cân bằng mà chúng ta có thể kham nổi, trên cơ sở chi phí cho lưu trữ và xử lý đã giảm mạnh.

Pat Helland, một trong những chuyên gia hàng đầu thế giới về thiết kế cơ sở dữ liệu, mô tả sự thay đổi cơ bản này trong một bài báo có tựa đề "Nếu bạn có quá nhiều dữ liệu, thì 'đủ tốt' là đủ tốt" ("if You Have Too Much Data, Then 'Good enough' is Good enough."). Sau khi xác định một số nguyên tắc cốt lõi của thiết kế truyền thống mà nay đã bị xói mòn bởi dữ liệu lộn xộn với nguồn gốc và độ chính xác khác nhau, ông đưa ra các hệ quả: "Chúng ta không còn có thể giả vờ rằng mình đang sống trong một thế giới sạch". Việc xử lý dữ liệu lớn đòi hỏi một sự mất mát thông tin không thể tránh khỏi - Helland gọi đó là "tổn hao". Nhưng bù lại, nó cho ra một kết quả nhanh chóng. "Nếu chúng ta bị tổn hao một số câu trả lời cũng không sao - đó vẫn luôn là những gì việc kinh doanh cần", Helland kết luận.

Thiết kế cơ sở dữ liệu truyền thống hứa hẹn sẽ cung cấp những kết quả luôn luôn nhất quán. Ví dụ nếu yêu cầu số dư tài khoản ngân hàng, bạn trông đợi sẽ nhận được con số chính xác. Và nếu yêu cầu nó một vài giây sau đó, bạn muốn hệ thống đưa ra cùng kết quả, với giả thuyết là không có thay đổi gì. Tuy nhiên, khi lượng dữ liệu thu thập phát triển và lượng người truy cập hệ thống tăng lên thì việc duy trì sự nhất quán này trở nên khó khăn hơn.

Các bộ dữ liệu lớn không tồn tại ở một nơi, chúng có xu hướng được phân bổ trên nhiều ổ đĩa cứng và máy tính. Để đảm bảo độ tin cậy và tốc độ, một bản ghi có thể được lưu trữ ở hai hoặc ba địa điểm khác nhau. Nếu bạn cập nhật bản ghi tại một địa điểm, dữ liệu ở các địa điểm khác sẽ không còn đúng nữa cho đến khi bạn cũng cập nhật nó. Trong khi các hệ thống truyền thống có một độ trễ để thực hiện tất cả các cập nhật, thì điều này không thực tế với dữ liệu được phân bổ rộng rãi và máy chủ phải bận

rộn với hàng chục ngàn truy vấn mỗi giây. Khi đó, việc chấp nhận tính hỗn độn chính là một dạng giải pháp.

Sự thay đổi này được đặc trưng bởi sự phổ biến của Hadoop, một đối thủ mã nguồn mở của hệ thống MapReduce của Google, rất tốt khi xử lý những lượng lớn dữ liệu. Nó thực hiện điều này bằng cách chia dữ liệu thành những phần nhỏ hơn và chia chúng ra cho các máy khác. Vì dự kiến phần cứng sẽ hỏng hóc, nên nó tạo ra sự dư thừa. Nó đặt giả thuyết dữ liệu không được sạch sẽ và trật tự - trong thực tế, nó cho rằng dữ liệu là quá lớn để được làm sạch trước khi xử lý. Mặc dù việc phân tích dữ liệu điển hình đòi hỏi một chuỗi thao tác được gọi là "trích xuất, chuyển giao, và tải", hoặc ETL (extract, transfer, and load) để chuyển dữ liệu đến nơi nó sẽ được phân tích, Hadoop bỏ qua những chi tiết như vậy. Thay vào đó, nó nghiễm nhiên chấp nhận rằng lượng dữ liệu là quá lớn nên không thể di chuyển và phải được phân tích ngay tại chỗ.

Đầu ra của Hadoop không chính xác bằng của các cơ sở dữ liệu quan hệ: nó không đáng tin để có thể dùng cho việc khởi động một con tàu vũ trụ hoặc xác nhận các chi tiết tài khoản ngân hàng. Nhưng đối với nhiều công việc ít quan trọng hơn, khi một câu trả lời cực kỳ chính xác là không cần thiết, thì nó thực hiện thủ thuật nhanh hơn rất nhiều so với các hệ thống khác. Hãy nghĩ tới những công việc như phân chia một danh sách khách hàng để gửi tới một số người một chiến dịch tiếp thị đặc biệt. Sử dụng Hadoop, công ty thẻ tín dụng Visa đã có thể giảm thời gian xử lý hồ sơ kiểm tra của hai năm, khoảng 73 tỷ giao dịch, từ một tháng xuống chỉ còn 13 phút. Việc tăng tốc xử lý như vậy là mang tính đột phá đối với các doanh nghiệp.

Kinh nghiệm của ZestFinance, một công ty được thành lập bởi cựu giám đốc thông tin của Google, Douglas Merrill, nhấn mạnh điểm này. Công nghệ của nó giúp người cho vay quyết định có hay không cung cấp những khoản vay ngắn hạn tương đối nhỏ

cho những người có vẻ như có điểm tín dụng kém. Tuy nhiên, trong khi điểm tín dụng truyền thống là chỉ dựa trên một số ít tín hiệu mạnh như các thanh toán chậm trước đây, thì ZestFinance phân tích một số lượng lớn các biến "yếu kém". Trong năm 2012, nó đã tự hào đưa ra một tỷ giá mặc định cho các khoản vay, một phần ba ít hơn so với mức trung bình trong ngành. Nhưng cách duy nhất để làm cho hệ thống hoạt động là chấp nhận sự hỗn độn.

"Một trong những điều thú vị", Merrill nói, "là không có ai mà tất cả các trường thông tin đều được điền đủ. Luôn luôn có một số lượng lớn dữ liệu bị thiếu". Ma trận thông tin do ZestFinance tập hợp là vô cùng tản mạn, một tập tin cơ sở dữ liệu đầy ắp những trường bị thiếu. Vì vậy, công ty "quy trách nhiệm" cho các dữ liệu bị thiếu. Ví dụ khoảng 10 phần trăm khách hàng của ZestFinance được liệt kê là đã chết - nhưng hóa ra điều đó chẳng ảnh hưởng đến việc trả nợ. "Vì vậy, rõ ràng là khi chuẩn bị hủy diệt những thây ma, hầu hết mọi người cho rằng không có khoản nợ nào sẽ được hoàn trả. Nhưng từ dữ liệu của chúng tôi, có vẻ như các thây ma đều trả lại khoản vay của mình", Merrill lém lỉnh kể tiếp.

Đổi lại việc sống chung với sự hỗn độn, chúng ta có được những dịch vụ rất có giá trị, những thứ lẽ ra không thể có ở phạm vi và quy mô của chúng với những phương pháp và công cụ truyền thống. Theo một số ước tính thì chỉ 5 phần trăm của tất cả dữ liệu kỹ thuật số là "có cấu trúc" - nghĩa là ở dạng thích hợp để đưa vào một cơ sở dữ liệu truyền thống. Nếu không chấp nhận sự hỗn độn thì 95 phần trăm còn lại của dữ liệu phi cấu trúc, chẳng hạn các trang web và phim, sẽ hoàn toàn ở trong bóng tối. Bằng cách cho phép sự không chính xác, chúng ta mở cửa vào một thế giới đầy những hiểu biết chưa được khai thác.

Xã hội đã thực hiện hai sự đánh đổi ngấm ngầm đã trở nên quen thuộc trong cách chúng ta ứng xử đến nỗi ta thậm chí không

xem chúng như những sự đánh đổi, mà chỉ như trạng thái tự nhiên của sự vật. Thứ nhất, chúng ta cho rằng mình không thể sử dụng được thật nhiều dữ liệu, vì vậy chúng ta không sử dụng. Nhưng sự hạn chế đó ngày càng mất đi ý nghĩa, và có rất nhiều thứ có thể đạt được nếu sử dụng một cái gì đó tiệm cận N = tất cả.

Sự đánh đổi thứ hai là về chất lượng của thông tin. Trong kỷ nguyên của dữ liệu nhỏ, khi chúng ta chỉ thu thập được một ít thông tin thì tính chính xác của nó phải là cao nhất có thể. Điều đó hợp lý. Trong nhiều trường hợp, điều này vẫn còn cần thiết. Nhưng đối với nhiều thứ khác, sự chính xác nghiêm ngặt ít quan trọng hơn việc nắm bắt được nhanh chóng những nét đại cương hay bước tiến triển theo thời gian của chúng.

Cách chúng ta nghĩ về việc sử dụng toàn bộ các thông tin so với những mảnh nhỏ của nó, và cách chúng ta có thể đi đến đánh giá cao sự lỏng lẻo thay vì tính chính xác, sẽ có những ảnh hưởng sâu sắc lên tương tác của chúng ta với thế giới. Khi kỹ thuật dữ-liệu-lớn trở thành một phần thường lệ của cuộc sống hàng ngày, chúng ta với tư cách một xã hội có thể bắt đầu cố gắng hiểu thế giới từ một góc nhìn lớn hơn, toàn diện hơn nhiều so với trước đây, một kiểu N = tất cả. Chúng ta có thể chấp nhận vết mờ và sự không rõ ràng trong những lĩnh vực mà mình vẫn thường đòi hỏi sự rõ ràng và chắc chắn, ngay cả khi chúng chỉ là một sự rõ ràng giả tạo và một sự chắc chắn không hoàn hảo. Chúng ta có thể chấp nhận điều này với điều kiện đổi lại chúng ta có được một hiểu biết hoàn chỉnh hơn về thực tại - tương đương với một bức tranh trừu tượng, trong đó từng nét vẽ là lộn xộn nếu được xem xét thật gần, nhưng khi bước lùi lại, ta có thể thấy một bức tranh hùng vĩ.

Dữ liệu lớn, với sự nhấn mạnh vào các bộ dữ liệu toàn diện và sự hỗn độn, giúp chúng ta tiến gần hơn tới thực tế so với sự phụ thuộc vào dữ liệu nhỏ và độ chính xác. Sự hấp dẫn của “một số” và “chắc chắn” là điều dễ hiểu. Hiểu biết của chúng ta về thế giới

có thể đã không đầy đủ và đôi khi sai lầm khi chúng ta bị hạn chế trong những gì chúng ta có thể phân tích, nhưng có một điều khá chắc chắn là nó mang lại một sự ổn định đáng yên tâm. Bên cạnh đó, vì bị kìm hãm trong dữ liệu có thể thu thập và khảo sát, chúng ta đã không phải đối mặt với sự cưỡng bách để có được tất cả mọi thứ, để xem tất cả mọi thứ từ mọi góc độ có thể. Và trong giới hạn hẹp của dữ liệu nhỏ, chúng ta vẫn không có được bức tranh lớn hơn dù có thể tự hào về độ chính xác của mình - thậm chí bằng cách đo các chi tiết vụn vặt đến một phần n độ.

Rốt cuộc, dữ liệu lớn có thể đòi hỏi *chúng ta* thay đổi, để trở nên thoải mái hơn với sự rối loạn và sự không chắc chắn. Các cấu trúc của sự chính xác, dù dường như cho chúng ta những ý nghĩa trong cuộc sống - kiểu như cái cọc tròn phải chui vào cái lỗ tròn; rằng chỉ có một câu trả lời cho một câu hỏi - lại dễ bị bóp méo hơn so với mức độ chúng ta có thể thừa nhận. Tuy nhiên sự thừa nhận, thậm chí đón nhận, tính linh hoạt này sẽ đưa chúng ta đến gần hơn với thực tế.

Những thay đổi trong tư duy này là những chuyển đổi căn bản, chúng dẫn tới một sự thay đổi thứ ba có khả năng phá hủy một tập quán còn cơ bản hơn của xã hội: ý tưởng về việc hiểu được các lý do đằng sau tất cả những gì xảy ra. Thay vào đó, như chương tiếp theo sẽ giải thích, việc tìm được các mối liên kết trong dữ liệu và hành động dựa trên chúng thường có thể là đủ tốt rồi.

4. TƯƠNG QUAN

Greg Linden ở tuổi 24 khi tạm nghỉ chương trình nghiên cứu tiến sĩ về trí tuệ nhân tạo tại Đại học Washington vào năm 1997, để làm việc tại một công ty bán sách trực tuyến của địa phương mới vừa thành lập.

Công ty mới hoạt động hai năm nhưng đã là một doanh nghiệp phát đạt. “Tôi yêu thích ý tưởng bán sách và bán kiến thức - và giúp đỡ mọi người tìm thấy mẩu kiến thức kế tiếp mà họ muốn thưởng thức”, ông hồi tưởng. Cửa hàng đó là Amazon.com, và họ thuê Linden với tư cách một kỹ sư phần mềm để đảm bảo cho trang web chạy trơn tru.

Amazon không chỉ có những chuyên gia kỹ thuật là nhân viên của mình. Vào thời điểm đó, họ cũng thuê hơn một chục nhà phê bình và biên tập viên để viết đánh giá và giới thiệu những cuốn sách mới. Mặc dù câu chuyện của Amazon quen thuộc với nhiều người, nhưng ít người còn nhớ nội dung của nó đã được chế tác bởi bàn tay con người. Các biên tập viên và nhà phê bình đánh giá và lựa chọn các cuốn sách được đưa lên các trang web của Amazon. Họ chịu trách nhiệm cho những gì được gọi là “Tiếng nói Amazon” - một trong những món trang sức vương giả và một nguồn lợi thế cạnh tranh của công ty. Một bài báo trên *Wall StreetJournal* vào thời điểm đó đã đón nhận chúng như những bài phê bình sách có ảnh hưởng nhất của quốc gia, bởi chúng đã tạo được rất nhiều doanh thu.

Sau đó, Jeff Bezos, người sáng lập và Giám đốc điều hành của Amazon, bắt đầu thử nghiệm một ý tưởng có sức thuyết phục mạnh mẽ: Liệu công ty có thể giới thiệu những cuốn sách cụ thể cho khách hàng dựa trên những sở thích mua sắm riêng biệt của họ? Ngay từ đầu, Amazon đã thu được nhiều dữ liệu về các khách hàng: những gì họ mua, những cuốn sách nào họ chỉ nhìn nhưng không mua, và họ nhìn chúng bao lâu. Những cuốn sách nào họ

đã mua cùng nhau.

Số lượng dữ liệu là rất lớn tới mức lúc đầu Amazon đã xử lý nó theo cách thông thường: lấy một mẫu và phân tích nó để tìm những điểm tương đồng giữa các khách hàng. Các khuyến nghị đưa ra là khá thô thiển. Mua một cuốn sách về Ba Lan và bạn sẽ bị "dội bom" bằng các thông tin về giá vé đi Đông Âu. Mua một cuốn về trẻ sơ sinh và bạn sẽ bị ngập với thông tin về những cuốn tương tự. "Họ có xu hướng mời bạn những biến thể nho nhỏ của những gì bạn mua trước đó, đến vô cùng tận", James Marcus, một nhà phê bình sách của Amazon từ 1996 đến 2001, nhớ lại trong hồi ký của mình mang tên *Amazonia.* "Cứ như bạn đi mua sắm với thằng đần vậy".

Greg Linden nhìn thấy một giải pháp. Ông nhận ra rằng hệ thống khuyến nghị không nhất thiết phải so sánh khách hàng với những người khác, một công việc phức tạp về kỹ thuật. Tất cả những gì cần thiết là tìm ra những liên kết giữa chính các sản phẩm với nhau. Năm 1998 Linden và đồng nghiệp của ông đã đăng ký một bằng sáng chế về kỹ thuật lọc cộng tác được gọi là "item-to-item" ("từ-mục-đến-mục"). Sự thay đổi trong cách tiếp cận đã làm nên một sự khác biệt lớn.

Bởi các tính toán có thể được thực hiện trước thời hạn, nên các khuyên nghị là rất nhanh. Phương pháp này cũng linh hoạt, có khả năng làm việc trên nhiều loại sản phẩm. Vì vậy, khi Amazon chuyển sang bán các mặt hàng khác với sách, hệ thống cũng có thể đề xuất phim hoặc lò nướng bánh. Và các khuyến nghị là tốt hơn nhiều so với trước vì hệ thống sử dụng tất cả các dữ liệu. "Câu nói đùa trong nhóm là nếu nó làm việc hoàn hảo thì Amazon sẽ chỉ giới thiệu cho bạn một cuốn sách mà thôi - đó là cuốn kế tiếp mà bạn sẽ mua", Linden nhớ lại.

Bây giờ công ty phải quyết định những gì sẽ xuất hiện trên trang web. Nội dung do máy tạo ra như những kiến nghị cá nhân và

những danh sách bán chạy nhất, hoặc những đánh giá được viết bởi các biên tập viên của Amazon? Những gì người xem nói, hoặc những gì các nhà phê bình nói? Đó là một trận chiến của chuột và người.

Khi Amazon làm một thử nghiệm so sánh doanh thu nhờ các biên tập viên với doanh thu nhờ các nội dung do máy tính tạo ra, kết quả thậm chí khá khác nhau. Nội dung máy tạo ra từ dữ liệu đã mang lại doanh thu cao hơn rất nhiều. Máy tính có thể không biết tại sao khách hàng đọc Ernest Hemingway cũng có thể muốn mua F. Scott Fitzgerald. Nhưng điều đó dường như không quan trọng. Tiếng leng keng của máy tính tiền mới quan trọng. Cuối cùng các biên tập viên đã chứng kiến tỷ lệ doanh số dựa trên những đánh giá trực tuyến của họ, và nhóm bị giải tán.

“Tôi rất buồn vì đội ngũ biên tập viên đã thất bại”, Linden nhớ lại. “Tuy nhiên, số liệu không nói dối, và chi phí thì rất cao”.

Ngày nay một phần ba doanh số bán hàng của Amazon được cho là kết quả của các hệ thống giới thiệu và cá nhân hóa. Với những hệ thống này, Amazon đã khiến nhiều đối thủ cạnh tranh bị phá sản: không chỉ những hiệu sách và cửa hàng âm nhạc lớn, mà cả những nhà bán sách địa phương nghĩ rằng mối liên hệ cá nhân của họ sẽ bảo vệ được họ khỏi những cơn gió của sự thay đổi. Thực tế, công việc của Linden đã cách mạng hóa thương mại điện tử, khi phương pháp này được áp dụng bởi gần như tất cả mọi người. Với NetAix, một công ty cho thuê phim trực tuyến, ba phần tư đơn đặt hàng mới đến từ các khuyến nghị. Theo sự dẫn đường của Amazon, hàng ngàn trang web có thể giới thiệu sản phẩm, nội dung, bạn bè và các nhóm mà không cần biết lý do mọi người lại có thể sẽ quan tâm đến chúng.

Biết *tại sao* có thể là thú vị, nhưng nó không quan trọng để kích thích bán hàng. Biết *cái gì* mới cuốn hút những cú nhấp chuột. Sự hiểu biết này có sức mạnh để định hình lại nhiều ngành công

nghiệp, chứ không chỉ thương mại điện tử. Nhân viên bán hàng trong tất cả các lĩnh vực từ lâu đã được cho biết rằng cần phải hiểu những gì làm cho khách hàng quan tâm, để nắm bắt những lý do đằng sau các quyết định của họ. Những kỹ năng chuyên nghiệp và nhiều năm kinh nghiệm đã được đánh giá cao. Dữ liệu lớn cho thấy có một phương pháp tiếp cận khác, trong một số góc độ là thực dụng hơn. Các hệ thống khuyến nghị sáng tạo của Amazon đưa ra được những mối tương quan có giá trị mà không cần biết nguyên nhân phía sau. Biết *cái gì,* chứ không phải *tại sao*, là đủ tốt rồi.

Những dự đoán và sở thích

Các mối tương quan cũng có ích trong một thế giới dữ-liệu-nhỏ, nhưng trong bối cảnh của dữ-liệu-lớn thì phải nói là chúng thực sự nổi bật. Thông qua chúng, người ta có thể thu được hiểu biết một cách dễ dàng hơn, nhanh hơn và rõ ràng hơn trước đây.

Tại cốt lõi của nó, một mối tương quan sẽ định lượng mối quan hệ thống kê giữa hai giá trị dữ liệu. Một tương quan mạnh có nghĩa là khi một giá trị dữ liệu thay đổi, thì giá trị dữ liệu kia rất có khả năng cũng thay đổi. Chúng ta đã thấy mối tương quan mạnh như vậy với Xu hướng Dịch cúm của Google: càng nhiều người trong một khu vực địa lý tìm kiếm những từ khóa cụ thể qua Google, thì càng có nhiều người tại khu vực đó mắc bệnh cúm. Ngược lại, một mối tương quan yếu có nghĩa là khi một giá trị dữ liệu thay đổi, ảnh hưởng của nó tới giá trị dữ liệu kia là nhỏ. Ví dụ chúng ta có thể kiểm tra tương quan giữa chiều dài mái tóc của các cá nhân và mức độ hạnh phúc của họ, và nhận thấy chiều dài mái tóc tỏ ra không có tác dụng gì đặc biệt trong việc cho chúng ta biết về mức độ hạnh phúc.

Các mối tương quan cho phép chúng ta phân tích một hiện tượng không phải bằng việc làm sáng tỏ hoạt động bên trong của nó, mà bằng cách xác định một phương tiện đo lường hữu ích

cho nó. Tất nhiên, ngay cả các mối tương quan mạnh cũng không bao giờ hoàn hảo. Rất có thể chúng hành xử tương tự chỉ vì sự ngẫu nhiên. Chúng ta có thể chỉ đơn giản "bị lừa bởi sự ngẫu nhiên", như một câu của nhà kinh nghiệm luận Nassim Nicholas Taleb. Với tương quan thì không có sự chắc chắn, mà chỉ có xác suất. Nhưng nếu một mối tương quan là mạnh thì khả năng của một liên kết sẽ cao. Nhiều khách hàng của Amazon có thể chứng thực điều này bằng cách chỉ vào một kệ sách đầy các khuyến nghị của công ty.

Bằng cách xác định một phương tiện đo lường thực sự tốt cho một hiện tượng, các mối tương quan giúp chúng ta nắm bắt được hiện tại và dự đoán được tương lai: nếu A thường xảy ra cùng với B, chúng ta cần phải xem chừng B để dự đoán rằng A sẽ xảy ra. Sử dụng B như một phương tiện đo lường sẽ giúp chúng ta nắm bắt những gì có thể xảy ra cùng với A, ngay cả khi chúng ta không thể đo lường hoặc quan sát được A một cách trực tiếp. Quan trọng hơn, nó cũng giúp chúng ta dự đoán những gì có thể xảy ra với A trong tương lai. Tất nhiên, các mối tương quan không thể nói trước tương lai, chúng chỉ có thể dự đoán nó với một xác suất nhất định. Nhưng khả năng đó là cực kỳ có giá trị.

Hãy xem trường hợp của Walmart. Đó là nhà bán lẻ lớn nhất thế giới, với hơn hai triệu nhân viên và doanh thu hàng năm khoảng 450 tỷ đôla- một khoản tiền lớn hơn GDP của bốn phần năm các nước trên thế giới. Trước khi web đưa ra quá nhiều dữ liệu thì có lẽ Walmart giữ tập dữ liệu lớn nhất của các công ty Mỹ. Trong những năm 1990 nó đã cách mạng hóa ngành bán lẻ bằng cách ghi lại tất cả sản phẩm như là dữ liệu thông qua một hệ thống được gọi là Liên kết Bán lẻ (Retail Link). Điều này cho phép các nhà cung cấp của Walmart theo dõi tỷ lệ và khối lượng bán hàng và hàng tồn kho. Việc tạo ra sự rõ ràng của thông tin này đã giúp công ty buộc các nhà cung cấp phải tự lo việc lưu trữ của họ. Trong nhiều trường hợp Walmart không tiếp nhận "quyền sở hữu" của một sản phẩm cho đến khi nó được bán, do đó loại bỏ

rủi ro hàng tồn kho và giảm được chi phí. Walmart sử dụng dữ liệu để thực sự trở thành cửa hàng ủy thác lớn nhất thế giới.

Các dữ liệu lịch sử có thể cho thấy những gì nếu chúng được phân tích một cách đúng đắn? Walmart đã làm việc với các chuyên gia phân tích số liệu từ Teradata, trước đây là công ty uy tín National Cash Register, để khám phá những mối tương quan thú vị. Năm 2004 Walmart cẩn thận xem xét cơ sở dữ liệu khổng lồ các giao dịch trong quá khứ của nó: mỗi khách hàng mua những mặt hàng gì và tổng chi phí, có những gì khác ở trong giỏ hàng, thời gian trong ngày, thậm chí cả thời tiết. Bằng cách đó, công ty nhận thấy rằng trước một cơn bão, không chỉ doanh số bán hàng của đèn pin tăng, mà cả mức bán Pop-Tarts, một món ăn sáng có đường của Mỹ, cũng tăng. Vì vậy, khi những cơn bão sắp đến, Walmart xếp những hộp Pop-Tarts ở ngay phía trước cửa hàng, bên cạnh các đồ tiếp tế bão, để tăng sự tiện lợi cho khách hàng - và tăng mạnh doanh số.

Trong quá khứ, một ai đó tại công ty sẽ cần có linh cảm trước để thu thập dữ liệu và thử nghiệm ý tưởng. Bây giờ, bởi có quá nhiều dữ liệu và những công cụ tốt hơn, các mối tương quan có thể được phát hiện một cách nhanh chóng hơn và ít tốn kém. (Nhưng cần nói rõ rằng chúng ta phải thận trọng: khi số lượng các điểm dữ liệu tăng với cấp độ lớn, chúng ta cũng thấy nhiều mối tương quan giả mạo hơn - những hiện tượng có vẻ như có mối liên hệ ngay cả khi chúng không phải như vậy. Điều này đòi hỏi chúng ta phải lưu tâm nhiều hơn, vì chúng ta chỉ mới bắt đầu đánh giá nó.)

Từ lâu trước khi có dữ liệu lớn, việc phân tích mối tương quan đã chứng tỏ là có giá trị. Khái niệm này được Ngài Francis Galton, người anh em họ của Charles Darwin, đưa ra vào năm 1888 sau khi ông nhận thấy một mối quan hệ giữa chiều cao và chiều dài cánh tay của những người đàn ông. Tính toán học đằng sau nó là tương đối đơn giản và chắc chắn - đó hóa ra là một trong những

đặc tính quan trọng, và đã giúp làm cho nó trở thành một trong những phép đo thống kê được sử dụng rộng rãi. Tuy nhiên, trước dữ liệu lớn, tính hữu dụng của nó bị hạn chế. Vì dữ liệu khan hiếm và việc thu thập tốn kém, nên các nhà thống kê thường chọn một phương tiện đo lường thay thế, sau đó thu thập các dữ liệu có liên quan và thực hiện phân tích tương quan để tìm hiểu xem phương tiện đó tốt tới đâu. Nhưng làm thế nào để chọn phương tiện đúng?

Để hướng dẫn họ, các chuyên gia sử dụng những giả thuyết dựa trên các lý thuyết - những ý tưởng trừu tượng về phương thức hoạt động của sự vật. Dựa trên những giả thuyết như vậy, họ thu thập dữ liệu và sử dụng phân tích tương quan để xác minh xem các phương tiện thay thế có phù hợp không. Nếu chúng không phù hợp, sau đó các nhà nghiên cứu thường cố gắng kiên định thực hiện lại, vì biết đâu các dữ liệu đã bị thu thập một cách sai lầm. Nếu thất bại thì cuối cùng họ mới phải thừa nhận rằng giả thuyết, hoặc thậm chí lý thuyết nền tảng của nó, còn thiếu sót và phải được sửa đổi. Kiến thức phát triển thông qua quá trình thử-và-sai như thế. Và nó diễn ra quá chậm, vì những thành kiến cá nhân và tập thể đã che mờ những giả thuyết chúng ta phát triển, chúng ta áp dụng chúng như thế nào, và do đó những phương tiện thay thế mà chúng ta đã chọn. Đó là một quá trình phức tạp, nhưng khả thi trong một thế giới dữ-liệu-nhỏ.

Trong thời đại dữ-liệu-lớn, việc ra quyết định để khảo sát những biến nào bằng cách chỉ dựa trên các giả thuyết sẽ không còn hiệu quả nữa. Các bộ dữ liệu là quá lớn và lĩnh vực được xem xét có lẽ quá phức tạp. May mắn thay, nhiều trong số những hạn chế vốn trói buộc chúng ta vào một cách tiếp cận dựa-trên-giả-thuyết đã không còn tồn tại với cùng mức độ như vậy nữa. Chúng ta bây giờ có quá nhiều dữ liệu dễ tiếp cận và khả năng tính toán tới mức không cần phải chăm chỉ chọn một hoặc một số ít phương tiện đo lường thay thế và khảo sát từng cái. Việc phân tích điện toán tinh vi bây giờ có thể xác định được phương tiện tối ưu -

như nó đã làm cho Xu hướng Dịch cúm của Google, sau khi "cày" qua gần nửa tỷ mô hình toán học.

Chúng ta không còn nhất thiết phải đòi hỏi một giả thuyết chuyên môn về một hiện tượng để bắt đầu hiểu thế giới của mình. Vì vậy, chúng ta không cần phát triển một khái niệm về những gì mọi người tìm kiếm khi nào và ở nơi nào bệnh cúm lây lan. Chúng ta không cần có một ý niệm mơ hồ về cách các hãng hàng không định giá vé của họ. Chúng ta không cần quan tâm đến thị hiếu của người mua hàng Walmart. Thay vào đó chúng ta có thể đặt dữ liệu lớn vào trong phép phân tích tương quan, để rồi nó sẽ cho chúng ta biết những câu hỏi tìm kiếm nào là các phương tiện đo lường tốt nhất cho bệnh cúm, liệu giá vé máy bay có khả năng tăng, hoặc những gì các công dân đang lo lắng chuẩn bị tránh bão sẽ muốn sử dụng. Thay cho việc tiếp cận dựa-trên-giả-thuyết, chúng ta có thể sử dụng cách tiếp cận dựa-trên-dữ-liệu. Các kết quả của chúng ta có thể ít bị chi phối và chính xác hơn, và chúng ta sẽ gần như chắc chắn nhận được chúng nhanh hơn nhiều.

Việc dự đoán dựa trên các mối tương quan chính là hạt nhân của dữ liệu lớn. Các phân tích tương quan bây giờ được sử dụng thường xuyên tới mức đôi khi chúng ta không còn đánh giá nổi mức độ xâm nhập của chúng nữa. Và việc ứng dụng này sẽ tăng. Ví dụ điểm tín dụng tài chính đang được sử dụng để dự đoán hành vi cá nhân. Công ty Fair Isaac Corporation, bây giờ được gọi là FICO, phát minh điểm tín dụng trong những năm cuối thập niên 1950. Năm 2011 FICO còn thiết lập "Điểm Ghi Nhớ Dùng Thuốc". Để xác định khả năng người ta sẽ dùng thuốc đến mức nào, FICO phân tích một loạt các biến - bao gồm cả những biến có vẻ không liên quan, chẳng hạn như họ đã sống bao lâu tại cùng địa chỉ, họ có kết hôn không, họ đã làm bao lâu với cùng một công việc, họ có sở hữu một chiếc xe hơi không. Điểm số ước lượng sẽ giúp các nhà cung cấp dịch vụ y tế tiết kiệm được tiền bằng cách cho họ biết những bệnh nhân nào cần được nhắc nhở.

Không có gì là quan hệ nhân quả giữa việc sở hữu xe hơi và uống thuốc kháng sinh theo chỉ dẫn; liên kết giữa chúng là tương quan thuần túy. Nhưng những kết quả như vậy cũng đủ để giám đốc điều hành của FICO mạnh miệng tuyên bố trong năm 2011: "Chúng tôi biết những gì bạn sẽ làm vào ngày mai đấy".

Những nhà môi giới dữ liệu khác đang thâm nhập vào cuộc chơi tương quan, như được phản ảnh trong loạt bài mang tính tiên phong "What They Know" ("Những Điều Họ Biết") của *Wall Street Journal*. Experian có một sản phẩm được gọi là Hiểu Thấu Thu Nhập để ước tính mức thu nhập của người dân mà một phần dựa trên cơ sở lịch sử tín dụng của họ. Nó phát triển điểm số bằng cách phân tích cơ sở dữ liệu lịch sử tín dụng khổng lồ của nó đối với dữ liệu thuế ẩn danh từ Sở Thuế Vụ Hoa Kỳ. Doanh nghiệp phải chi khoảng $10 để xác nhận thu nhập của một người thông qua các biểu khai thuế, trong khi Experian bán ước tính của nó ít hơn $1. Vì vậy, trong những trường hợp như thế này, việc sử dụng phương tiện đo lường thay thế sẽ hiệu quả hơn là đi hàn huyên để có được những điều thực tế. Tương tự, một văn phòng tín dụng khác, Equifax, bán một "Chỉ số Khả năng trả tiền" và một "Chỉ số Chi tiêu tùy ý" hứa hẹn dự đoán được sự tình trạng đầy hay vơi của ví tiền cá nhân.

Việc sử dụng các mối tương quan đang được mở rộng hơn nữa. Aviva, một công ty bảo hiểm lớn, đã nghiên cứu ý tưởng sử dụng các báo cáo tín dụng và dữ liệu tiếp thị người tiêu dùng như những phương tiện đo lường để phân tích mẫu máu và nước tiểu cho các ứng viên nhất định. Mục đích là để xác định những người có thể có nguy cơ cao mắc các bệnh như huyết áp cao, tiểu đường, hoặc trầm cảm. Phương pháp này sử dụng dữ liệu về lối sống bao gồm hàng trăm biến như các sở thích, các trang web truy cập, và mức độ xem truyền hình, cũng như ước tính thu nhập của họ. Mô hình dự đoán Aviva, được phát triển bởi Deloitte Consulting, được xem là thành công trong việc xác định nguy cơ sức khỏe. Những công ty bảo hiểm khác như Prudential

và AIG đã xem xét các sáng kiến tương tự. Lợi ích là nó có thể cho phép người nộp đơn xin bảo hiểm tránh được việc phải cung cấp mẫu máu và nước tiểu, mà chẳng ai thích, và các công ty bảo hiểm lại phải trả tiền cho việc đó. Chi phí xét nghiệm khoảng $125 cho mỗi người, trong khi các phương pháp tiếp cận hoàn toàn dựa-trên-dữ-liệu chỉ tốn khoảng $5.

Với một số người, phương pháp này nghe có vẻ đáng sợ, bởi vì nó dựa trên những hành vi dường như không mấy liên quan với nhau. Nó giống như việc các công ty có thể ẩn danh để làm gián điệp mạng, theo dõi từng cú nhấp chuột. Mọi người có thể sẽ cân nhắc kỹ lưỡng trước khi xem những trang web của các môn thể thao cực đoan hay xem hài kịch tôn vinh sự trầm cảm nếu họ cảm thấy điều này có thể dẫn đến phí bảo hiểm cao hơn. Phải thừa nhận rằng việc cản trở tự do của người dân trong tương tác với thông tin sẽ là điều tệ hại. Nhưng mặt khác, lợi ích trong việc khiến bảo hiểm dễ dàng hơn và ít tốn kém hơn sẽ mang lại kết quả là có nhiều người tham gia bảo hiểm hơn, đó là một điều tốt cho xã hội, chưa kể cũng tốt cho các công ty bảo hiểm.

Tuy nhiên, sản phẩm "đinh" của các mối tương quan dữ-liệu-lớn chính là cửa hàng bán lẻ giảm giá Target của Mỹ, đã có nhiều năm sử dụng các dự đoán dựa trên các mối tương quan dữ-liệu-lớn. Trong một phóng sự đặc biệt, Charles Duhigg, một phóng viên kinh doanh của *New York Times*, kể lại cách Target biết được một người phụ nữ đã có thai mà thậm chí chẳng cần người mẹ tương lai phải nói ra. Về cơ bản, phương pháp của họ là khai thác dữ liệu và để cho các mối tương quan làm công việc của chúng.

Việc biết nếu một khách hàng có thể mang thai là rất quan trọng cho các nhà bán lẻ, vì mang thai là một thời điểm bước ngoặt cho các cặp vợ chồng, khi hành vi mua sắm của họ sẽ sẵn sàng thay đổi. Họ có thể bắt đầu đi tới những cửa hàng mới và phát triển những sở thích thương hiệu mới. Những nhà tiếp thị của Target tìm đến bộ phận phân tích để xem có cách nào phát hiện ra

những khách hàng mang thai thông qua mô hình mua sắm của họ.

Nhóm phân tích xem xét lại lịch sử mua sắm của những phụ nữ đăng ký quà cho trẻ sơ sinh. Họ nhận thấy những phụ nữ này mua rất nhiều kem dưỡng da không mùi vào khoảng tháng thứ ba của thai kỳ, và vài tuần sau đó, họ thường mua những chất bổ trợ như magiê, canxi, và kẽm. Cuối cùng, nhóm phát hiện khoảng hai mươi sản phẩm, được sử dụng như các phương tiện đo lường, cho phép công ty tính toán được một loại "điểm dự đoán mang thai" cho từng khách hàng thanh toán bằng thẻ tín dụng hoặc sử dụng thẻ cửa hàng hoặc phiếu khuyến mãi. Các mối tương quan thậm chí cho phép nhà bán lẻ ước tính được thời hạn sinh con trong một khoảng hẹp, do vậy họ có thể gửi những phiếu khuyến mãi thích hợp cho từng giai đoạn của thai kỳ. Quả đúng với cái tên của doanh nghiệp này, "Target", nghĩa là "Mục tiêu".

Trong cuốn sách *The Power of Hahit (Sức mạnh của Thói quen)*, tác giả Duhigg kể tiếp câu chuyện này. Vào một ngày nọ, một người đàn ông giận dữ xông vào một cửa hàng Target ở Minnesota để gặp người quản lý. "Con gái tôi nhận được cái này trong thùng thư!", ông ta hét lên. "Con bé vẫn còn đang học trung học, vậy mà ông gửi phiếu khuyến mãi mua quần áo và giường cũi trẻ sơ sinh? ông đang khuyến khích con tôi có thai hả?". Thế nhưng khi người quản lý gọi lại cho ông ta một vài ngày sau đó để xin lỗi, ông ta lại tỏ ra hòa nhã và thậm chí chính ông ta phải xin lỗi người quản lý.



Đoạn phim tác giả Duhigg giải thích và minh họa câu chuyện

Việc tìm kiếm các phương tiện đo lường thay thế trong các bối cảnh xã hội chỉ là một trong nhiều cách tận dụng các kỹ thuật liên quan đến dữ-liệu-lớn. Bên cạnh đó, các mối tương quan với các kiểu dữ liệu mới để giải quyết các nhu cầu hàng ngày cũng tỏ ra mạnh mẽ không kém là.

Một trong số đó là phương pháp phân tích dự đoán, bắt đầu được sử dụng rộng rãi trong kinh doanh để dự đoán các sự kiện trước khi chúng xảy ra. Thuật ngữ này có thể được dùng để chỉ một thuật toán giúp phát hiện một ca khúc nổi tiếng, thường được sử dụng trong ngành công nghiệp âm nhạc để cung cấp cho các hãng ghi âm một ý tưởng tốt hơn về nơi để họ đầu tư. Kỹ thuật này cũng được sử dụng để ngăn chặn những hỏng hóc lớn về cơ khí hoặc cấu trúc: đặt các cảm biến trên máy móc, động cơ, hoặc

cơ sở hạ tầng để có thể theo dõi các mô hình dữ liệu mà chúng phát ra, chẳng hạn như nhiệt độ, độ rung, độ căng, và âm thanh, và để phát hiện những thay đổi có thể dự báo trước các sự cố.

Khái niệm nền tảng của phương pháp trên là khi sự vật hỏng, chúng thường không hỏng tất cả cùng một lúc, mà dần dần theo thời gian. Khi được trang bị dữ liệu cảm biến, việc phân tích tương quan và các phương pháp tương tự có thể xác định các mô hình cụ thể, các dấu hiệu, thường nảy sinh trước khi một cái gì đó hỏng - tiếng nổ của động cơ, nhiệt độ quá cao từ một động cơ, và những thứ tương tự. Từ đó, ta chỉ cần tìm kiếm mô hình để biết khi nào một cái gì đó tỏ ra bất ổn. Việc phát hiện sự bất thường sớm cho phép hệ thống gửi một cảnh báo để có thể thay một bộ phận mới hoặc chỉnh sửa sai sót trước khi sự cố thực sự xảy ra. Mục đích là để xác định một phương tiện đo lường tốt, sau đó quan sát nó, và qua đó dự đoán các sự kiện trong tương lai.

Công ty vận chuyển UPS đã sử dụng các phân tích dự đoán từ cuối những năm 2000 để theo dõi đội xe 60 ngàn chiếc tại Hoa Kỳ và biết khi nào cần thực hiện bảo dưỡng phòng ngừa. Mọi sự cố trên đường đều có thể khiến phải hủy bỏ hay trì hoãn việc giao và nhận hàng. Vì vậy, để phòng ngừa, UPS thường thay thế một số bộ phận sau hai hoặc ba năm. Nhưng điều đó không hiệu quả, vì một số bộ phận vẫn còn tốt. Từ khi chuyển sang phân tích dự báo, công ty đã tiết kiệm được hàng triệu đôla bằng cách đo và giám sát các bộ phận riêng lẻ và thay thế chúng chỉ khi cần thiết. Trong một trường hợp, dữ liệu thậm chí tiết lộ rằng toàn bộ một nhóm các xe mới có một bộ phận bị khiếm khuyết có thể gây rắc rối, trừ khi được phát hiện trước khi đưa vào sử dụng.

Tương tự như vậy, các bộ cảm biến được gắn vào cầu và các tòa nhà để theo dõi các dấu hiệu hao mòn. Chúng cũng được sử dụng trong các nhà máy hóa chất lớn và các nhà máy lọc dầu, những nơi mà nếu một bộ phận bị hỏng có thể làm ngưng trệ sản xuất.

Chi phí cho việc thu thập và phân tích dữ liệu để biết khi nào phải hành động sớm là thấp hơn so với chi phí của việc ngừng sản xuất.

Lưu ý rằng các phân tích dự đoán có thể không giải thích nguyên nhân của một vấn đề; nó chỉ cho thấy có vấn đề tồn tại. Nó sẽ cảnh báo bạn rằng một động cơ quá nóng, nhưng nó có thể không cho bạn biết tình trạng đó là do một đai quạt bị sờn hay do một nắp đậy không được vặn chặt. Các mối tương quan cho biết *cái gì,* nhưng không cho biết *tại sao,* tuy nhiên như chúng ta đã thấy, biết *cái gì* thường là đủ tốt rồi.

Phương pháp tương tự đang được áp dụng trong y tế, để ngăn ngừa các "hỏng hóc" của cơ thể con người. Khi bệnh viện gắn một mớ ống, dây điện, và các dụng cụ cho bệnh nhân, một dòng lớn dữ liệu được tạo ra. Chỉ riêng điện tâm đồ đã ghi 1.000 thông số mỗi giây. Tuy nhiên, đáng chú ý là chỉ có một phần nhỏ của dữ liệu hiện đang được sử dụng hoặc lưu giữ. Hầu hết dữ liệu bị bỏ đi, ngay cả khi nó có thể giữ những đầu mối quan trọng về tình trạng và phản ứng với phương pháp điều trị của bệnh nhân. Nếu được giữ lại và tổng hợp với dữ liệu của các bệnh nhân khác, chúng có thể tiết lộ những thông tin đặc biệt về việc phương pháp điều trị nào có khả năng tiến triển tốt và phương pháp nào thì không.

Việc vứt bỏ dữ liệu có thể thích hợp khi chi phí và độ phức tạp của việc thu thập, lưu trữ và phân tích nó là cao, nhưng bây giờ điều này không còn đúng nữa. Tiến sĩ Carolyn McGregor và một nhóm các nhà nghiên cứu tại Viện Công nghệ của Đại học Ontario và IBM đã làm việc với một số bệnh viện để xây dựng phần mềm giúp các bác sĩ ra các quyết định chẩn đoán tốt hơn khi chăm sóc trẻ sinh non. Phần mềm thu nhận và xử lý dữ liệu bệnh nhân trong thời gian thực, theo dõi 16 dòng dữ liệu khác nhau, chẳng hạn như nhịp tim, nhịp thở, nhiệt độ, huyết áp, và mức oxy trong máu, tổng cộng tới khoảng 1.260 điểm dữ liệu

mỗi giây.

Hệ thống có thể phát hiện những thay đổi tinh tế về tình trạng của trẻ thiếu tháng, có thể báo hiệu tình trạng nhiễm trùng 24 giờ trước khi các triệu chứng rõ ràng xuất hiện. “Bạn không thể nhìn thấy nó bằng mắt thường, nhưng một máy tính thì có thể”, tiến sĩ McGregor giải thích. Hệ thống không dựa vào quan hệ nhân quả mà dựa vào các mối tương quan. Nó cho biết *cái gì*, không cho biết *tại sao.* Nhưng điều đó đáp ứng được mục đích của nó. Việc cảnh báo trước cho phép các bác sĩ điều trị nhiễm trùng sớm với những bước can thiệp y tế nhẹ nhàng hơn, hoặc cảnh báo họ sớm hơn nếu việc điều trị tỏ ra không hiệu quả. Điều này cải thiện tình trạng của bệnh nhân. Kỹ thuật nói trên rất đáng được áp dụng cho thật nhiều bệnh nhân hơn và trong nhiều điều kiện hơn. Thuật toán tự nó có thể không đưa ra các quyết định, nhưng máy đang làm những gì máy làm tốt nhất, để giúp những người chăm sóc làm những gì họ làm tốt nhất.

Điều đáng chú ý là việc phân tích dữ-liệu-lớn của tiến sĩ McGregor đã có thể xác định những mối tương quan mà theo một nghĩa nào đó đối nghịch với hiểu biết thông thường của các bác sĩ. Ví dụ bà phát hiện rằng các dấu hiệu sống rất ổn định thường được phát hiện trước khi bị nhiễm trùng nghiêm trọng. Đây là điều kỳ lạ, vì chúng ta cứ tưởng rằng sự xuống cấp của các cơ quan duy trì sự sống sẽ xảy ra trước một đợt nhiễm trùng toàn diện. Người ta có thể hình dung ra cảnh hàng thế hệ các bác sĩ kết thúc ngày làm việc của họ bằng cách liếc nhìn một bệnh án bên cạnh giường, thấy các dấu hiệu sống ổn định của trẻ sơ sinh, và an tâm đi về nhà - để rồi nhận được một cuộc gọi hoảng loạn từ phòng y tá trực vào lúc nửa đêm thông báo rằng điều cực kỳ bi thảm đã xảy ra và bản năng của họ đã được đặt không đúng chỗ.

Dữ liệu của McGregor cho thấy rằng sự ổn định của các trẻ thiếu tháng, thay vì là một dấu hiệu của sự cải thiện, lại giống như sự bình lặng trước con bão - cứ như cơ thể của trẻ sơ sinh nói cho

các cơ quan nhỏ xíu của mình hãy sẵn sàng cho điều tệ hại sắp xảy ra. Chúng ta không thể biết chắc chắn, vì những gì dữ liệu cho thấy là một tương quan, chứ không phải quan hệ nhân quả. Nhưng chúng ta biết rằng nó đòi hỏi các phương pháp thống kê được áp dụng cho một lượng lớn các dữ liệu để tiết lộ sự liên hợp ẩn này. Nếu có ai còn nghi ngờ thì đây: dữ liệu lớn cứu được nhiều mạng sống.

Ảo tưởng và sự soi sáng

Trong một thế giới dữ-liệu-nhỏ, vì có rất ít dữ liệu, nên cả những nghiên cứu về nguyên nhân lẫn phân tích tương quan đều bắt đầu với một giả thuyết, sau đó được kiểm nghiệm để hoặc thấy sai hoặc xác minh. Nhưng vì cả hai phương pháp đòi hỏi một giả thuyết để bắt đầu, nên cả hai đều nhạy cảm với thành kiến và trực giác sai lầm. Và các dữ liệu cần thiết thường không có sẵn. Ngày nay, với rất nhiều dữ liệu xung quanh và nhiều hơn nữa sẽ tới, những giả thuyết như vậy không còn quan trọng đối với phân tích tương quan.

Có một sự khác biệt mới đang dần trở nên quan trọng. Trước thời dữ liệu lớn, một phần do sức mạnh tính toán không đầy đủ, nên phần lớn việc phân tích tương quan sử dụng những tập hợp lớn dữ liệu bị giới hạn vào việc tìm kiếm các mối quan hệ tuyến tính. Trong thực tế, tất nhiên, nhiều mối quan hệ là phức tạp hơn nhiều. Với những phân tích tinh vi hơn, chúng ta có thể xác định được những mối quan hệ phi tuyến tính trong dữ liệu.

Ví dụ: trong nhiều năm các nhà kinh tế và các nhà khoa học chính trị tin rằng hạnh phúc và thu nhập có tương quan trực tiếp - tăng thu nhập và một người trung bình sẽ được hạnh phúc hơn. Tuy nhiên việc quan sát dữ liệu trên một biểu đồ cho thấy một tình trạng phức tạp hơn đã diễn ra. Đối với các mức thu nhập dưới một ngưỡng nhất định, mỗi sự gia tăng trong thu nhập dẫn tới sự gia tăng đáng kể trong hạnh phúc, nhưng trên mức đó thì

việc tăng thu nhập hầu như không cải thiện được hạnh phúc của một cá nhân. Nếu ta thể hiện điều này trên đồ thị, đường biểu diễn sẽ là một đường cong thay vì một đường thẳng như giả định bằng phân tích tuyến tính.

Phát hiện này rất quan trọng cho các nhà hoạch định chính sách. Nếu nó là một mối quan hệ tuyến tính thì việc nâng cao thu nhập của tất cả mọi người nhằm cải thiện hạnh phúc chung sẽ có ý nghĩa. Nhưng một khi mối liên hệ phi tuyến tính đã được xác định thì lời tư vấn sẽ chuyển thành tập trung vào việc tăng thu nhập cho người nghèo, vì dữ liệu cho thấy điều này sẽ mang lại nhiều hiệu quả cho đồng tiền.

Chuyện sẽ trở nên phức tạp hơn nhiều, chẳng hạn như khi mối quan hệ tương quan là nhiều mặt hơn. Ví dụ các nhà nghiên cứu tại Đại học Harvard và MIT đã khảo sát sự chênh lệch của việc chủng ngừa bệnh sởi trong dân cư - một số nhóm được chủng ngừa trong khi những nhóm khác thì không. Đầu tiên sự chênh lệch này dường như tương quan với số tiền người dân chi cho chăm sóc sức khỏe. Tuy nhiên, việc xem xét kỹ hơn cho thấy mối tương quan không phải một đường gọn gàng mà là một đường cong kỳ quặc. Khi mọi người chi tiêu nhiều hơn cho chăm sóc sức khỏe, sự chênh lệch về tiêm chủng giảm xuống (như có thể được dự kiến), nhưng khi họ chi tiêu nhiều hơn nữa, điều đáng ngạc nhiên là nó lại tăng lên - một số người rất giàu dường như né tránh tiêm chủng ngừa sởi. Thông tin này rất quan trọng với các viên chức y tế công, nhưng phân tích tương quan tuyến tính đơn giản sẽ không thể phát hiện được nó.

Các chuyên gia bây giờ đang phát triển các công cụ cần thiết để xác định và so sánh các mối tương quan phi tuyến tính. Đồng thời, các kỹ thuật phân tích tương quan đang được hỗ trợ và tăng cường bởi một tập hợp phát triển nhanh chóng các phương pháp tiếp cận và phần mềm mới mẻ, có thể rút ra được những mối liên hệ phi-nhân-quả trong dữ liệu từ nhiều góc độ khác nhau - giống

như cách các họa sĩ lập thể đã cố gắng nắm bắt được hình ảnh khuôn mặt của một người phụ nữ từ nhiều góc độ cùng một lúc. Một trong những phương pháp mới và mạnh nhất có thể được tìm thấy trong lĩnh vực đang phát triển của phân tích mạng lưới. Nó cho phép lập bản đồ, đo lường và tính toán các nút và các liên kết cho tất cả mọi thứ từ bạn bè của một người trên Facebook, tới những phán quyết tòa án nào trích dẫn những tiền lệ nào, hoặc ai gọi ai trên điện thoại di động của họ. Cùng với nhau, những công cụ này giúp trả lời những câu hỏi thực nghiệm phi-quan-hệ-nhân-quả.

Cuối cùng, trong thời đại của dữ liệu lớn, các kiểu phân tích mới này sẽ dẫn đến một làn sóng hiểu biết mới và dự đoán hữu ích. Chúng ta sẽ thấy những liên kết chưa bao giờ thấy trước đó. Chúng ta sẽ nắm bắt được những động lực kỹ thuật và xã hội phức tạp từ lâu đã trốn tránh nhận thức của chúng ta mặc cho những nỗ lực tốt nhất. Nhưng quan trọng nhất, các phân tích phi-quan-hệ-nhân-quả này sẽ cải thiện sự hiểu biết của chúng ta về thế giới bằng cách chủ yếu hỏi *cái gì* chứ không hỏi *tại sao.*

Lúc đầu, điều này nghe có vẻ khác thường. Nhưng xét cho cùng, là con người, chúng ta mong muốn hiểu biết thế giới thông qua các liên kết nhân quả; chúng ta muốn tin rằng mỗi hiệu ứng đều có một nguyên nhân, chỉ cần chúng ta nhìn đủ kỹ lưỡng. Liệu có nên xem việc biết được các lý do nền tảng của thế giới là khát vọng cao nhất của chúng ta?

Để chắc chắn, đã có một cuộc tranh luận triết học hàng thế kỷ trước đây về việc liệu có tồn tại quan hệ nhân quả. Nếu mỗi thứ đều là do một cái gì đó khác gây ra, thì logic sẽ ra lệnh rằng chúng ta không được tự do để quyết định bất cứ điều gì cả. Ý chí của con người sẽ không tồn tại, vì mọi quyết định mà chúng ta đưa ra và mọi suy nghĩ chúng ta hình thành đều do cái gì khác gây ra, trong khi bản thân nó cũng là hệ quả của một nguyên nhân khác, cứ như thế... Quỹ đạo của cả cuộc sống sẽ chỉ đơn

giản được xác định bởi các nguyên nhân dẫn đến các hiệu ứng. Do đó các triết gia đã tranh cãi về vai trò của quan hệ nhân quả trong thế giới của chúng ta, và đôi khi nó chống lại ý chí tự do. Tuy nhiên cuộc tranh luận trừu tượng kể trên không phải là những gì chúng ta theo đuổi ở đây.

Thay vào đó, khi chúng ta nói rằng con người nhìn thế giới qua các quan hệ nhân quả, chúng ta đang đề cập đến hai cách thức cơ bản con người giải thích và hiểu thế giới: thông qua mối quan hệ nhân quả nhanh chóng, phi thực tế; hoặc thông qua cách thực nghiệm chậm rãi, theo phương pháp quan hệ nhân quả. Dữ liệu lớn sẽ làm thay đổi vai trò của cả hai.

Cách thứ nhất chính là mong ước trực giác của chúng ta muốn thấy các kết nối nhân quả. Chúng ta mang định kiến phải giả định các nguyên nhân, ngay cả khi chúng không tồn tại.

Điều này không phải do văn hóa, sự dạy dỗ hay mức độ học vấn. Thay vào đó, nghiên cứu cho thấy nó là một vấn đề liên quan đến sự vận hành của nhận thức con người. Khi thấy hai sự kiện xảy ra, cái này tiếp sau cái kia, tâm trí của chúng ta bị thôi thúc phải nhìn thấy chúng trong những quan hệ nhân quả.

Hãy xem ba câu sau đây: “Cha mẹ của Fred đến muộn. Những người cung cấp thực phẩm phải đến sớm. Fred đã tức giận”. Khi đọc xong, chúng ta ngay lập tức trực cảm vì sao Fred tức giận - không phải vì những người cung cấp thực phẩm phải đến sớm, mà vì cha mẹ anh ta đến muộn. Thật ra, chúng ta không có cách nào biết được điều này từ các thông tin được cung cấp. Ấy vậy mà tâm trí của chúng ta vẫn khăng khăng tạo ra những điều chúng ta giả định là mạch lạc, những quan hệ nhân quả từ các dữ kiện hạn chế đó.

Daniel Kahneman, một giáo sư tâm lý học tại Princeton từng đoạt giải Nobel kinh tế năm 2002, sử dụng ví dụ này để đưa ra

giả thuyết rằng chúng ta có hai phương thức suy nghĩ. Một là nhanh chóng và mất ít nỗ lực, cho phép chúng ta đi đến kết luận trong vài giây. Phương thức khác là chậm và khó khăn, đòi hỏi chúng ta phải suy nghĩ một vấn đề cụ thể. Cách suy nghĩ nhanh chóng mang rất nhiều định kiến để hướng tới việc "nhìn ra" những liên kết nhân quả ngay cả khi chúng không tồn tại. Nó tạo nên định kiến để xác nhận kiến thức và niềm tin sẵn có của chúng ta. Trong lịch sử cổ đại, cách suy nghĩ nhanh này đã giúp chúng ta sống sót qua một môi trường nguy hiểm, trong đó chúng ta thường phải quyết định một cách nhanh chóng với thông tin hạn chế. Nhưng nó thường không đủ để thiết lập nguyên nhân thực sự của một hiệu ứng.

Theo Kahneman, thật không may rằng não của chúng ta rất thường xuyên lười suy nghĩ một cách chậm rãi và có phương pháp. Thay vào đó, chúng ta cho phép cách suy nghĩ nhanh chóng thắng thế. Kết quả là chúng ta thường "nhìn ra" những quan hệ nhân quả tưởng tượng, và do đó về cơ bản đã hiểu sai thế giới.

Cha mẹ thường nói với con cái của họ rằng chúng mắc bệnh cúm vì không đội mũ hoặc mang găng tay trong thời tiết lạnh. Tuy nhiên, không hề có quan hệ nhân quả trực tiếp giữa đội mũ, mang găng và mắc bệnh cúm. Nếu chúng ta ghé một nhà hàng và sau đó bị bệnh thì một cách trực giác, chúng ta đổ lỗi cho các thực phẩm chúng ta ăn ở đó (và có thể sẽ tránh nhà hàng này trong tương lai), mặc dù thực phẩm có thể không liên quan gì với bệnh tật của chúng ta. Dạ dày của chúng ta có thể nhiễm khuẩn qua nhiều cách, chẳng hạn như bắt tay một người bị nhiễm bệnh. Bên tư-duy-nhanh của bộ não chúng ta được lập trình sẵn để nhanh chóng nhảy tới bất cứ kết luận nhân quả nào nó có thể thấy. Do đó, nó thường dẫn chúng ta đến những quyết định sai lầm.

Trái ngược với suy nghĩ thông thường, trực giác của con người

về mối quan hệ nhân quả như vậy không giúp hiểu biết của chúng ta về thế giới sâu sắc thêm. Trong nhiều trường hợp, nó chỉ hơn một chút so với một đường tắt nhận thức, cho chúng ta ảo tưởng về cái nhìn sâu sắc, nhưng trong thực tế lại bỏ roi chúng ta trong bóng tối về thế giới xung quanh. Cũng giống như việc lấy mẫu là một đường tắt chúng ta sử dụng vì không thể xử lý được tất cả các dữ liệu, nhận thức về quan hệ nhân quả là một đường tắt mà não chúng ta sử dụng để tránh phải suy nghĩ khó khăn và chậm chạp.

Trong một thế giới dữ-liệu-nhỏ, việc chỉ ra những kiểu trực giác nhân quả sai như thế nào cần một thời gian dài. Điều này sẽ thay đổi. Trong tương lai, các mối tương quan dữ-liệu-lớn sẽ thường xuyên được sử dụng để bác bỏ trực giác quan hệ nhân quả của chúng ta, cho thấy thường có rất ít, nếu như có, liên kết thống kê giữa kết quả và nguyên nhân giả định của nó.

Có lẽ bài học đó sẽ làm cho chúng ta suy nghĩ nghiêm khắc hơn (và chậm rãi hơn) khi muốn hiểu thế giới. Nhưng ngay cả suy nghĩ chậm rãi của chúng ta - cách thứ hai tìm ra những quan hệ nhân quả - cũng sẽ thấy vai trò của nó được biến đổi bởi các mối tương quan dữ liệu lớn.

Trong cuộc sống hàng ngày, chúng ta tư duy theo kiểu nhân quả nhiều đến nỗi có thể tin rằng quan hệ nhân quả sẽ dễ dàng được chỉ ra. Thực tế khó khăn hơn nhiều. Không giống với các mối tương quan, trong đó tính chất toán học là khá rõ ràng, đối với quan hệ nhân quả thì chẳng có phương thức toán học nào rõ ràng để “chứng minh” cả.

Chúng ta thậm chí không thể diễn tả các mối quan hệ nhân quả một cách dễ dàng trong những phương trình chuẩn. Do đó ngay cả nếu ta suy nghĩ chậm rãi và nghiêm khắc, việc tìm kiếm các mối quan hệ nhân quả cuối cùng vẫn rất khó khăn. Bởi vì tâm trí của chúng ta quen với một thế giới nghèo thông tin, chúng ta bị

cám dỗ phải lý giải với ít dữ liệu, mặc dù rất thường xuyên có quá nhiều yếu tố cùng phối hợp làm giảm ảnh hưởng đến một nguyên nhân cụ thể.

Hãy xem trường hợp của vắc-xin chống bệnh dại. Vào ngày 6 tháng 7 năm 1885, nhà hóa học người Pháp Louis Pasteur được giới thiệu với cậu bé chín tuổi Joseph Meister, vốn bị một con chó dại tấn công dã man. Trước đó Pasteur đã phát minh ra việc chủng ngừa và đang nghiên cứu một loại vắc-xin thử nghiệm chống bệnh dại. Cha mẹ của Meister xin Pasteur sử dụng vắc-xin để điều trị con trai của họ. Ông đồng ý, và Joseph Meister đã sống sót. Trên báo chí, Pasteur trở nên nổi tiếng vì đã cứu cậu bé khỏi một cái chết chắc chắn và đau đớn.

Nhưng có phải ông đã làm việc đó? Thực ra, trung bình chỉ có một trong bảy người bị chó dại cắn là mắc bệnh. Thậm chí giả sử vắc-xin thử nghiệm của Pasteur có hiệu quả, thì đã có tới khoảng 85 phần trăm khả năng là cậu bé sẽ sống sót.

Trong ví dụ này, việc dùng vắc-xin được xem là đã chữa khỏi cho Joseph Meister. Nhưng có hai mối quan hệ nhân quả được đặt ra: thứ nhất là giữa vắc-xin và vi-rút bệnh dại, thứ hai là giữa việc bị một con cắn và việc phát triển bệnh. Ngay cả khi điều thứ nhất đúng, điều thứ hai chỉ đúng trong một số ít trường hợp.

Các nhà khoa học đã vượt qua thách thức này để chứng minh quan hệ nhân quả thông qua các thí nghiệm, trong đó nguyên nhân giả định có thể được chấp nhận hoặc loại bỏ một cách cẩn thận. Nếu những tác động xuất hiện tương ứng với việc nguyên nhân đó được áp dụng hay không thì nó cho thấy một mối quan hệ nhân quả. Càng kiểm soát các tình huống một cách cẩn thận thì khả năng mối quan hệ nhân quả mà bạn đã xác định là chính xác càng cao.

Do đó, giống như các mối tương quan, quan hệ nhân quả có thể

rất hiếm khi được chứng minh, mà chỉ xuất hiện với xác suất cao. Nhưng không giống như các mối tương quan, các thí nghiệm để suy ra các quan hệ nhân quả thường không thực tế hoặc tăng thách thức những vấn đề luân lý. Làm thế nào chúng ta có thể tiến hành một thử nghiệm quan hệ nhân quả để xác định lý do một số thuật ngữ tìm kiếm nhất định lại dự đoán được tốt nhất về bệnh cúm? Và với mũi tiêm bệnh dại, liệu chúng ta có đẩy hàng chục, có thể hàng trăm bệnh nhân - một phần trong “nhóm thực nghiệm,, không được tiêm - đến một cái chết đau đớn, mặc dù chúng ta đã có vắc-xin cho họ? Kể cả trường hợp thí nghiệm là thực tế, chúng vẫn còn đắt và tốn thời gian.

So với nó, việc phân tích phi-nhân-quả, chẳng hạn các mối tương quan, thường nhanh và rẻ tiền. Không giống như các liên kết nhân quả, chúng ta có các phương pháp toán học và thống kê để phân tích các mối quan hệ và các công cụ kỹ thuật số cần thiết để chứng minh sức mạnh của chúng với sự tự tin.

Hơn nữa, các mối tương quan không chỉ có giá trị cho riêng chúng mà còn vạch đường cho các cuộc điều tra nhân quả. Bàng cách cho chúng ta biết hai sự vật nào có khả năng liên kết, chúng cho phép ta tiếp tục điều tra xem một mối quan hệ nhân quả có tồn tại không, và nếu như vậy thì tại sao. Cơ chế lọc không tốn kém và nhanh này làm giảm chi phí của phân tích quan hệ nhân quả thông qua các thí nghiệm kiểm soát đặc biệt. Thông qua các mối tương quan chúng ta có thể có cái nhìn thoáng qua về các biến quan trọng mà sau đó chúng ta sử dụng trong các thí nghiệm để điều tra nguyên nhân.

Nhưng hãy cẩn trọng. Các mối tương quan là mạnh không chỉ vì chúng cung cấp những hiểu biết, mà còn vì những hiểu biết chúng cung cấp là tương đối rõ ràng. Những hiểu biết này thường bị che khuất khi chúng ta mang quan hệ nhân quả áp dụng vào tình huống. Ví dụ, Kaggle, một công ty tổ chức những cuộc thi khai-thác-dữ-liệu đã lập ra một cuộc thi vào năm 2012

về chất lượng của xe cũ, mở cho tất cả mọi người. Một đại lý xe cũ cung cấp dữ liệu cho các nhà thống kê tham gia cuộc thi để xây dựng một thuật toán nhằm dự đoán những chiếc xe bán đấu giá nào có khả năng gặp sự cố. Một phân tích tương quan cho thấy những chiếc xe sơn màu da cam ít bị khiếm khuyết hơn nhiều - khoảng một nửa tỷ lệ trung bình của các xe khác.

Ngay khi đọc điều này, chúng ta đã nghĩ tại sao nó lại có thể như vậy: Những người sở hữu xe hơi màu da cam có thể là những người đam mê xe hơi và chăm sóc xe của họ tốt hơn? Có phải với một màu đặc biệt nào đó, chiếc xe đã được sản xuất một cách cẩn thận hơn, được tinh chỉnh trong cả các khía cạnh khác nữa? Hoặc, có lẽ những chiếc xe màu cam là đáng chú ý hơn trên đường và do đó ít có khả năng bị tai nạn, vì vậy chúng ở trong tình trạng tốt hơn khi bán lại?

Chúng ta nhanh chóng bị vây hãm trong một lưới các giả thuyết nhân quả cạnh tranh với nhau. Nhưng những nỗ lực của chúng ta để làm sáng tỏ mọi việc theo cách này chỉ khiến cho chúng mờ mịt thêm. Các mối tương quan có tồn tại, chúng ta có thể biểu lộ chúng về mặt toán học. Nhưng chúng ta không thể dễ dàng làm điều tương tự cho các liên kết nhân quả. Vì vậy, chúng ta sẽ từ bỏ cố gắng giải thích lý do đằng sau các mối tương quan: *tại sao* thay vì *cái gì.* Nếu không, chúng ta có thể sẽ tư vấn cho những chủ sở hữu xe sơn những chiếc xe cũ của họ màu da cam để giúp cho máy ít bị hỏng - một suy nghĩ rất vô lý.

Trong những năm gần đây, các nhà khoa học đã cố gắng giảm chi phí thí nghiệm điều tra nhân quả, ví dụ bằng cách khéo léo kết hợp thêm các cuộc điều tra chọn mẫu thích hợp để tạo ra những cuộc “thử nghiệm giả”. Điều đó có thể giúp cho một số cuộc điều tra nhân quả trở nên dễ dàng hơn, nhưng vẫn khó lòng lấn át được lợi thế hiệu quả của các phương pháp phi-nhân-quả. hơn nữa, dữ liệu lớn tự nó hỗ trợ việc điều tra nhân quả vì nó hướng dẫn các chuyên gia hướng tới các nguyên nhân có triển

vọng để điều tra. Trong nhiều trường hợp, việc tìm kiếm sâu hơn cho quan hệ nhân quả sẽ diễn ra sau khi dữ liệu lớn đã thực hiện công việc của mình, khi chúng ta đặc biệt muốn điều tra *tại sao,* chứ không chỉ đánh giá cao vấn đề *cái gì.*

Quan hệ nhân quả sẽ không bị loại bỏ, nhưng nó không còn được tôn thờ như suối nguồn của hiểu biết nữa. Dữ liệu lớn đã truyền năng lượng cho các phân tích phi-nhân-quả để chúng thường xuyên thay thế các điều tra nhân quả. Câu hỏi hóc búa về vụ nổ các hố ga ở Manhattan là một ví dụ điển hình.

Cuộc chiến giữa người và hố

Mỗi năm vài trăm hố ga tại thành phố New York bắt đầu âm ỉ vì bên trong chúng bắt lửa. Đôi khi các nắp cống bằng gang, trọng lượng tới 300 pound (khoảng 150 kg), phát nổ, văng lên không trung cao tới mấy tầng nhà trước khi rơi xuống mặt đất. Đây chẳng phải chuyện hay ho gì.

Con Edison, công ty tiện ích công cộng cung cấp điện của thành phố, thực hiện kiểm tra và bảo trì thường xuyên các hố ga hàng năm. Trong quá khứ, về cơ bản nó dựa trên may rủi, hy vọng rằng một hố ga trong kế hoạch kiểm tra có thể là một trong số đang sẵn sàng nổ. Như vậy còn tốt hơn chút ít so với chỉ dạo bước kiểm tra ngẫu nhiên xuống Phố Wall. Năm 2007 Con Edison nhờ tới các nhà thống kê tại Đại học Columbia với hy vọng họ có thể sử dụng dữ liệu lịch sử của công ty về mạng lưới hố ga, chẳng hạn như những sự cố trước đây và cơ sở hạ tầng nào được kết nối với nhau, để dự đoán những hố ga nào có khả năng gặp sự cố, như vậy công ty sẽ biết được nơi tập trung nguồn lực của mình.

Đó là một vấn đề dữ-liệu-lớn phức tạp. Có 94.000 dặm cáp ngầm trong thành phố New York, đủ để quấn xung quanh Trái đất ba vòng rưỡi. Chỉ riêng Manhattan đã có khoảng 51.000 hố ga và tủ

điện. Một phần cơ sở hạ tầng này là từ thời Thomas Edison, thế nên công ty mới có tên như thế. Cứ 20 cáp thì có một đã được đặt trước năm 1930. Mặc dù hồ sơ được lưu giữ từ những năm 1880, nhưng chúng ở những dạng rất hỗn độn - và chưa bao giờ được tạo ra để phục vụ cho việc phân tích dữ liệu. Chúng đến từ bộ phận kế toán hoặc điều phối khẩn cấp nên được viết tay trên các "phiếu sự cố". Nếu chỉ nói rằng dữ liệu hỗn độn nghĩa là đã nói giảm một cách trắng trợn. Một ví dụ: các nhà thống kê tường trình rằng cái gọi là "tủ điện", một bộ phận phổ biến của cơ sở hạ tầng, có ít nhất 38 biến thể, chẳng hạn SB, S, S/B, S.B, S?B, S.B., SBX, S/BX, SB/X, S/XB, /SBX, S.BX, S&BX, S?BX, S BX, S/B/X, S BOX, SVBX, SERV BX, SERV-BOX, SERV/BOX, và SERVICE BOX. Một thuật toán máy tính phải hình dung ra tất cả những thứ đó.

"Các dữ liệu là vô cùng thô", người đứng đầu dự án Cynthia Rudin, nhà thống kê và khai thác dữ liệu, nay ở MIT, nhớ lại. "Tôi đã có một bản in của tất cả các bảng cáp khác nhau. Nếu mở ra, bạn thậm chí không thể giữ nó mà không bị rơi xuống sàn nhà. Và bạn phải tìm được ý nghĩa từ tất cả những thứ đó - để đào bới chúng lên mà tìm vàng, hoặc làm bất cứ điều gì để có được một mô hình dự đoán thực sự tốt".

Để làm việc, Rudin và nhóm của cô đã phải sử dụng tất cả các dữ liệu có sẵn, không chỉ là một mẫu, vì bất kỳ cái nào trong số hàng chục ngàn hố ga đều có thể là một quả bom nổ chậm đang đếm giờ. Vì vậy, nó nhất thiết hướng đến N = tất cả. Và mặc dù việc đưa ra được các lý lẽ mang tính nhân quả chắc hẳn rất hay ho, nhưng điều đó có thể cần cả một thế kỷ và kết quả vẫn sẽ sai hoặc không đầy đủ. Cách tốt hơn để thực hiện công việc là tìm các mối tương quan. Rudin ít quan tâm đến *tại sao* hơn *cái nào* - dù cô biết rằng khi nhóm ngồi đối diện các nhà điều hành của Con Edison, các chuyên viên thống kê phải biện minh cho cơ sở cách xếp thứ hạng của họ. Các dự đoán có thể được thực hiện bởi một cỗ máy, nhưng khách hàng lại là con người, và con người có xu hướng muốn tìm lý do, muốn hiểu.

Và việc khai thác dữ liệu làm lộ ra những thoi vàng mà Rudin hy vọng tìm thấy. Sau khi định dạng dữ liệu hỗn độn để máy tính có thể xử lý được, nhóm nghiên cứu bắt đầu với 106 dự đoán của một thảm họa hố ga lớn. Sau đó họ cô đọng danh sách cho một số ít các dấu hiệu mạnh nhất. Trong một thử nghiệm với mạng lưới điện của Bronx, họ đã phân tích tất cả các dữ liệu có trong tay, đến giữa năm 2008. Sau đó, họ sử dụng dữ liệu đó để dư đoán các điểm có vấn đề cho năm 2009. Nó đã đạt kết quả xuất sắc. Lần này nhóm 10 phần trăm hố ga nằm trên cùng trong danh sách của họ đã bao gồm tới 44 phần trăm các hố ga mà sau đó gặp sự cố nghiêm trọng.

Xét cho cùng, các yếu tố quan trọng nhất là tuổi của các dây cáp và liệu các hố ga đã trải qua những sự cố trước đó chưa. Những điều này hóa ra rất hữu ích, vì nó có nghĩa là dây cáp đồng của Con Edison có thể dễ dàng làm cơ sở cho việc xếp thứ hạng. Mà khoan. Tuổi và những sự cố trước đây sao? Chẳng phải chuyện đó hiển nhiên quá còn gì? Vâng, có và không. Một mặt, như nhà lý thuyết mạng Duncan Watts thường nói, “Một khi bạn đã biết câu trả lời thì mọi thứ đều tỏ ra hiển nhiên cả”. Nhưng mặt khác, điều quan trọng là phải nhớ rằng ngay từ đầu đã có tới 106 kiểu dự đoán trong mô hình. Việc đánh giá tầm quan trọng của chúng, sau đó xếp thứ tự ưu tiên cho hàng chục ngàn hố ga, mỗi hố với vô số biến đã tạo ra đến hàng triệu điểm dữ liệu, chưa kể bản thân dữ liệu không phải ở dạng có thể phân tích được. Chuyện này chẳng hề hiển nhiên hay rõ ràng.

Trường hợp những hố ga nổ nêu bật lên một điểm là dữ liệu đang được đưa vào sử dụng theo một cách mới để giải quyết các bài toán khó khăn trong thế-giới-thực. Tuy nhiên để đạt được điều này, chúng ta cần thay đổi cách làm việc. Chúng ta phải sử dụng tất cả các dữ liệu, nhiều nhất trong khả năng chúng ta có thể thu thập được, chứ không chỉ một phần nhỏ. Chúng ta phải chấp nhận sự hỗn độn thay vì xem sự chính xác như một ưu tiên hàng đầu. Và chúng ta phải đặt niềm tin của mình vào các mối tương

quan mà không cần hiểu biết đầy đủ về cơ sở quan hệ nhân quả cho các dự đoán.

Sự kết thúc của lý thuyết?

Dữ liệu lớn thay đổi cách thức chúng ta hiểu và khám phá thế giới. Trong thời đại của dữ liệu nhỏ, chúng ta được định hướng bởi các giả thuyết về cách thức thế giới hoạt động, để rồi sau đó chúng ta mới cố gắng xác nhận chúng bằng cách thu thập và phân tích dữ liệu. Trong tương lai, sự hiểu biết của chúng ta sẽ được dẫn dắt bởi sự phong phú của dữ liệu hơn là bởi các giả thuyết.

Những giả thuyết này thường được bắt nguồn từ các lý thuyết về tự nhiên hay các khoa học xã hội, những thứ này lại giúp giải thích và/hoặc dự đoán thế giới xung quanh chúng ta. Khi chuyển đổi từ một thế giới được điều khiển bởi giả thuyết sang một thế giới được điều khiển bởi dữ liệu, chúng ta có thể dại dột nghĩ rằng mình không còn cần các lý thuyết nữa.

Năm 2008, tổng biên tập của tạp chí *Wired* Chris Anderson đã loan báo rằng “cuộc đại hồng thủy dữ liệu sẽ khiến phương pháp khoa học trở nên lỗi thời”. Trong một bài viết được giới thiệu ngay trang bìa có tên là “The Petabyte Age” (“Thời đại Petabyte”), ông tuyên bố rằng nó dẫn tới “sự kết thúc của lý thuyết”. Quá trình truyền thống của khám phá khoa học - một giả thuyết được kiểm nghiệm trên thực tế bằng cách sử dụng một mô hình của các quan hệ nhân quả nền tảng - đang trên đà biến mất, Anderson kết luận. Nó sẽ được thay thế bằng phân tích thống kê của các mối tương quan thuần túy, phi lý thuyết.

Để hỗ trợ cho lập luận của mình, Anderson mô tả vật lý lượng tử đã trở thành một lĩnh vực gần như hoàn toàn lý thuyết, bởi vì các thí nghiệm là quá đắt, quá phức tạp, và quá lớn nên không mang tính khả thi. ông chỉ ra rằng có lý thuyết chẳng liên quan

gì với thực tế nữa. Để nêu ví dụ cho phương pháp mới, ông nhắc đến công cụ tìm kiếm Google và việc xác định trình tự gen. “Đây là thế giới mà những lượng lớn dữ liệu và môn toán học ứng dụng sẽ thay thế mọi công cụ khác”, ông viết. “Với đủ dữ liệu, các con số sẽ tự phát biểu cho chúng. Petabyte cho phép chúng ta khẳng định: “Tính tương quan là đủ”.”

Bài báo mở ra một cuộc tranh luận dữ dội và rất đáng quan tâm, mặc dù Anderson nhanh chóng rút lại tuyên bố táo bạo của mình. Nhưng lý lẽ của ông đáng để xem xét. Về bản chất, Anderson khẳng định rằng cho đến gần đây, khi muốn phân tích và hiểu thế giới xung quanh, chúng ta vẫn cần các lý thuyết để kiểm tra. Nhưng ngược lại, trong thời đại dữ-liệu-lớn, chúng ta không cần các lý thuyết nữa: chúng ta có thể chỉ nhìn vào dữ liệu. Nếu đúng như vậy, điều này sẽ cho thấy rằng tất cả các quy luật khái quát về cách thế giới hoạt động, cách con người cư xử, những gì người tiêu dùng mua, khi nào các bộ phận hỏng... đều có thể trở nên không thích hợp nữa khi bị thay thế bằng phân tích dữ liệu lớn.

“Sự kết thúc của lý thuyết” dường như ngụ ý rằng mặc dù các lý thuyết đã tồn tại trong các lĩnh vực chuyên môn như vật lý hay hóa học, việc phân tích dữ-liệu-lớn chẳng cần bất kỳ mô hình khái niệm nào. Điều này là phi lý.

Bản thân Dữ liệu lớn được hình thành dựa trên lý thuyết. Ví dụ, nó sử dụng các lý thuyết thống kê và toán học, và đôi khi sử dụng cả khoa học máy tính. Đúng, chúng không phải là những lý thuyết về động lực quan hệ nhân quả của một hiện tượng đặc biệt như trọng lực, nhưng dù sao chúng vẫn là những lý thuyết.

Và, như chúng ta đã chỉ ra, các mô hình dựa trên chúng có khả năng dự đoán rất hữu ích. Thật ra, dữ liệu lớn có thể cung cấp một cái nhìn tươi mát và những hiểu biết mới mẻ một cách chính xác vì nó không bị cản trở bởi lối suy nghĩ thông thường và

những thành kiến cố hữu tiềm ẩn trong các lý thuyết của một lĩnh vực cụ thể.

Hơn nữa, vì việc phân tích dữ-liệu-lớn được dựa trên các lý thuyết, ta không thể thoát khỏi chúng. Chúng định hình cả các phương pháp và các kết quả của chúng ta. Trước tiên là cách chúng ta lựa chọn dữ liệu. Các quyết định của chúng ta có thể được định hướng bởi sự tiện lợi: Phải chăng dữ liệu đã có sẵn? Hoặc bởi tính kinh tế: Liệu có thể thu thập được dữ liệu một cách ít tốn kém? Lựa chọn của chúng ta bị ảnh hưởng bởi các lý thuyết. Những gì chúng ta chọn sẽ ảnh hưởng tới những gì chúng ta tìm thấy, như các nhà nghiên cứu công nghệ số Danah Boyd và Kate Crawford đã lập luận. Xét cho cùng, Google đã sử dụng các từ khóa tìm kiếm như một phương tiện đo lường cho dịch cúm, chứ không sử dụng độ dài của tóc người. Tương tự như vậy, khi phân tích dữ liệu, chúng ta chọn những công cụ dựa trên các lý thuyết. Và khi giải thích kết quả, chúng ta lại áp dụng các lý thuyết. Thời đại của dữ liệu lớn rõ ràng không phải là không có lý thuyết - chúng có mặt khắp mọi nơi, với tất cả những gì chúng thừa hưởng.

Anderson xứng đáng được vinh danh khi nêu lên những câu hỏi xác đáng - và đặc biệt là ông đã làm thế sớm hơn những người khác. Dữ liệu lớn có thể không chỉ rõ vào "Sự kết thúc của lý thuyết", nhưng nó chuyển đổi một cách cơ bản cách chúng ta cảm nhận thế giới. Sự thay đổi này sẽ đòi hỏi rất nhiều công sức để làm quen. Nó thách thức nhiều tổ chức. Tuy nhiên, giá trị to lớn mà nó mang lại sẽ làm cho nó không chỉ là một sự đánh đổi đáng giá, mà còn là thứ không thể tránh khỏi.

Tuy nhiên trước khi đạt tới đó, cũng đáng để lưu tâm xem chúng ta đã tới đây như thế nào. Nhiều người trong ngành kỹ thuật cao muốn gán công trạng chuyển đổi cho các công cụ kỹ thuật số mới, từ các chip nhanh tới phần mềm hiệu quả, bởi vì họ là những người làm ra công cụ. Sự kỳ diệu của kỹ nghệ là quan

trọng, nhưng không quan trọng nhiều như người ta tưởng. Lý do sâu xa hơn của những xu hướng này là chúng ta có nhiều dữ liệu hơn rất nhiều. Và lý do chúng ta có nhiều dữ liệu hơn là vì chúng ta đã đưa nhiều khía cạnh hơn của thực tế vào một định dạng dữ liệu, cũng chính là chủ đề của chương kế tiếp.

5. DỮ LIỆU HÓA

Matthew Fontaine Maury là một sĩ quan Hải quân Hoa Kỳ đầy triển vọng. Trên đường nhận một nhiệm vụ mới tại Consort vào năm 1839, xe ngựa của ông đột nhiên trượt khỏi đường, lật nhào, và ném ông vào không khí. Ông bị ngã đau, gãy xương đùi và trẹo khớp gối. Khớp được một bác sĩ địa phương chỉnh lại vào vị trí, nhưng xương đùi thì được xếp rất tồi và vài ngày sau bị tháo ra để đặt lại. Những vết thương đã làm Maury, lúc đó mới 33 tuổi, bị liệt một phần và không còn thích hợp với biển. Sau gần ba năm hồi phục, Hải quân xếp cho ông công việc bàn giấy, phụ trách một nơi nghe chẳng hấp dẫn chút nào - Kho Bản đồ và Khí giới.

Hóa ra đó lại là nơi hoàn hảo cho ông. Là một hoa tiêu trẻ, Maury từng rất bực bội vì các con tàu cứ chạy ngoằn ngoèo trên đại dương thay vì đi theo những tuyến đường trực tiếp hơn. Khi ông hỏi các thuyền trưởng về chuyện này, họ trả lời rằng việc đi theo một tuyến đường quen thuộc sẽ tốt hơn là chấp nhận may rủi với một tuyến đường mình không nắm rõ bằng, vốn dĩ tiềm ẩn những nguy hiểm. Họ xem đại dương như là một địa hạt không thể đoán trước, nơi các thủy thủ phải đối mặt những bất ngờ với tất cả gió và sóng.

Tuy nhiên, từ những chuyến đi của ông, Maury biết rằng điều này không hoàn toàn đúng, ông nhìn ra những khuôn mẫu ở khắp mọi nơi. Trong một chặng dừng kéo dài tại Valparaiso, Chile, ông đã chứng kiến những cơn gió hoạt động chính xác cứ như đồng hồ. Một cơn gió mạnh vào chiều muộn sẽ đột nhiên dịu đi lúc mặt trời lặn và trở thành một làn gió nhẹ, cứ như thể ai đó vừa ngắt van. Trong một chuyên đi khác ông đã vượt qua dòng hải lưu xanh ấm áp Gulf Stream khi nó chảy giữa những khoảng tối của nước biển Đại Tây Dương. Trông nó thật khác biệt và ổn định, cứ như thể đó là dòng sông Mississippi vậy. Thật ra, người Bồ Đào Nha đã đi lại trên Đại Tây Dương hàng thế kỷ bằng cách

dựa vào các luồng gió đông và tây đều đặn được gọi là "gió mậu dịch".

Bất cứ khi nào chuẩn úy hải quân Maury đến một cảng mới, ông đều tìm kiếm những thuyền trưởng già để thu thập kiến thức của họ, dựa trên các trải nghiệm được truyền lại qua các thế hệ. Ông đã học được những kiến thức về thủy triều, gió, và hải lưu hoạt động theo quy luật, nhưng không hề được tìm thấy trong các sách và bản đồ mà Hải quân cấp cho các thủy thủ. Thay vào đó, họ dựa trên những bản đồ đôi khi cũ cả trăm năm, nhiều bản đồ có rất nhiều thiếu sót hoặc hoàn toàn không chính xác. Trong cương vị mới là người quản lý Kho Bản đồ và Quân dụng, ông tập trung khắc phục điều đó.

Khi nhận nhiệm vụ, ông kiểm kê các phong vũ biểu, la bàn, kính lục phân, và đồng hồ bấm giờ trong bộ sưu tập của kho. Ông cũng chú ý tới vô số những cuốn sách, bản đồ, và biểu đồ hàng hải có trong kho. Ông đã tìm thấy những thùng mốc đầy các sổ ghi chép cũ từ tất cả những chuyến đi trước đây của các thuyền trưởng Hải quân. Người tiền nhiệm của ông đã xem chúng là rác. Với những lời hài hước hoặc những hình phác thảo kỳ quặc trên lề các trang giấy, chúng đôi khi có vẻ giống như một cách để thoát khỏi sự nhàm chán của chuyến đi hơn là một sự ghi chép về hành trình của con tàu.

Nhưng khi Maury phủi bụi những cuốn sách ố màu nước biển và xem kỹ bên trong, ông thật sự thích thú. Đây là những thông tin ông cần: hồ sơ về gió, nước và thời tiết tại những địa điểm cụ thể trong những ngày cụ thể. Mặc dù một số bản ghi cung cấp được ít giá trị, nhiều bản khác đã cho thấy bạt ngàn thông tin hữu ích. Ghép tất cả chúng lại, Maury nhận thấy một hình thức hoàn toàn mới của biểu đồ điều hướng sẽ hoàn toàn khả thi. Maury và cả tá "máy tính" của ông - chức danh của những người tính toán số liệu - bắt đầu quá trình cần mẫn trích xuất và lập bảng các thông tin đã bị giam cầm bên trong các cuốn sổ ghi chép đang bị

hủy hoại.

Maury tổng hợp các dữ liệu và phân chia toàn bộ Đại Tây Dương thành các khối năm độ kinh tuyến và vĩ tuyến. Với từng phân khúc ông ghi nhiệt độ, tốc độ, hướng của gió và sóng, cùng với tháng, vì những điều kiện này khác nhau tùy thuộc vào thời gian trong năm. Khi kết hợp lại, dữ liệu cho thấy những mô hình và chỉ ra được những tuyến đường hiệu quả hơn.

Lời khuyên của nhiều thế hệ thủy thủ đôi khi đã đưa những con tàu thẳng tiến vào những vùng yên ả hoặc khiến chúng phải đọ sức với gió và dòng chảy ngược chiều. Trên một tuyến đường thông thường, từ New York đến Rio de Janeiro, các thủy thủ từ lâu đã có tư tưởng phải chống lại thiên nhiên thay vì dựa vào nó. Các hoa tiêu Mỹ được dạy tránh các nguy hiểm của một hành trình về phía nam thẳng đến Rio. Vì vậy, những con tàu của họ đã lướt theo dòng đông nam trước khi chuyển qua dòng tây nam sau khi vượt qua đường xích đạo. Khoảng cách đi thuyền thường lên tới ba lần xuyên suốt toàn bộ Đại Tây Dương. Tuyến đường phức tạp hóa ra lại là vô nghĩa. Một đường đơn giản trực tiếp về phía nam cũng đã là tốt.

Để tăng độ chính xác, Maury cần nhiều thông tin hơn. Ông đã tạo ra một phiếu chuẩn để ghi nhật ký dữ liệu của tàu và yêu cầu tất cả các tàu Hải quân Mỹ sử dụng và nộp lại khi kết thúc chuyến đi. Các tàu buôn rất muốn có được những sơ đồ của ông, nhưng Maury kiên quyết yêu cầu đổi lại họ phải nộp các phiếu ghi nhật ký tàu của họ (một phiên bản sớm của một mạng xã hội lan truyền). “Mỗi con tàu đi trên đại dương”, ông tuyên bố, “có thể từ nay về sau được xem như một đài quan sát nổi, một ngôi đền của khoa học”. Để tinh chỉnh các sơ đồ, ông đã tìm kiếm các điểm dữ liệu khác (giống như Google xây dựng trên thuật toán PageRank để bao gồm nhiều tín hiệu hơn). Ông yêu cầu các thuyền trưởng ném chai với các ghi chú cho thấy ngày, vị trí, gió, và dòng chảy phổ biến trên biển theo định kỳ, và vớt những chai như vậy khi

phát hiện ra chúng. Nhiều tàu cắm một lá cờ đặc biệt để cho thấy họ đã hợp tác với việc trao đổi thông tin (tiền thân của các biểu tượng chia sẻ liên kết sau này xuất hiện trên một số trang web).

Từ các dữ liệu, các tuyến đường biển tự nhiên đã tự thể hiện, nơi mà gió và dòng chảy là đặc biệt thuận lợi. Các sơ đồ của Maury cắt giảm được những hành trình dài, thường khoảng một phần ba, giúp các thương gia tiết kiệm được rất nhiều chi phí. “Cho đến khi có được những tài liệu của ông, tôi đã vượt qua đại dương trong mịt mù”, một thuyền trưởng đã viết lời tán thưởng như vậy. Và thậm chí cả những người đi biển sành sỏi, vẫn từ chối các sơ đồ mới lạ và dựa trên những cách truyền thống hoặc trực giác của họ, cũng đóng một vai trò hữu ích: nếu hành trình của họ mất nhiều thời gian hơn hoặc gặp thảm họa, xem như họ đã chứng minh tính tiện ích cho hệ thống của Maury. Đến năm 1855, khi xuất bản tác phẩm có uy tín *The Physical Geography of the Sea*, Maury đã vẽ được 1,2 triệu điểm dữ liệu. “Do đó, một thủy thủ trẻ, thay vì mò mẫm theo cách của mình cho đến khi ánh sáng của kinh nghiệm đến với anh ta... thì qua đây sẽ thấy rằng anh ta đã có kinh nghiệm của một ngàn hoa tiêu để hướng dẫn cho mình, cùng một lúc”, ông đã viết.

Công trình của ông có ý nghĩa quan trọng cho việc lắp đặt cáp điện báo xuyên Đại Tây Dương đầu tiên. Và, sau một vụ va chạm thảm khốc trên biển, ông đã nhanh chóng sắp đặt hệ thống các làn tàu vận chuyển mà ngày nay đã trở thành phổ biến. Thậm chí ông còn áp dụng phương pháp của mình cho thiên văn học: khi hành tinh Neptune được phát hiện vào năm 1846, Maury đã có ý tưởng tuyệt vời là phối hợp các tài liệu lưu trữ đã nhầm lẫn nhắc đến nó như một ngôi sao, và chúng đã giúp vẽ được quỹ đạo của Neptune. Maury đã hầu như bị bỏ qua trong các sách lịch sử Mỹ, có lẽ bởi con người gốc Virginia này đã từ chức khỏi Hải quân trong thời kỳ Nội chiến và phục vụ như một điệp viên ở Anh cho phe Liên minh. Nhưng nhiều năm trước đó, khi ông đến châu Âu để kêu gọi sự hỗ trợ quốc tế cho các sơ đồ của mình, bốn quốc gia

đã phong tước hiệp sĩ cho Maury, và ông đã nhận được huy chương vàng từ tám nước khác, bao gồm cả Vatican. Vào thời kỳ đầu của thế kỷ XXI, biểu đồ dẫn đường do Hải quân Mỹ xuất bản vẫn mang tên ông.

Trung tá Maury, "Thám tử của đại dương", là một trong những người đầu tiên nhận ra rằng có một thứ giá trị đặc biệt trong một gói tổng hợp rất lớn của dữ liệu, điều không thể có được với lượng dữ liệu nhỏ hơn - một nguyên lý cốt lõi của dữ liệu lớn. Về cơ bản, ông hiểu rằng những tập nhật ký hàng hải mốc meo của Hải quân đã thực sự tạo nên "dữ liệu" có thể khai thác, trích xuất và lập bảng. Khi làm như vậy, ông là một trong những người tiên phong của *dữ liệu hóa*, khai quật dữ liệu từ một nguồn mà không ai nghĩ rằng có chứa bất kỳ giá trị nào. Giống như Oren Etzioni tại Farecast, người đã sử dụng thông tin về giá cũ của ngành công nghiệp hàng không để tạo ra một công việc kinh doanh sinh lợi, hay các kỹ sư tại Google, những người đã tận dụng những câu hỏi tìm kiếm cũ để hiểu về sự lây lan của dịch cúm, Maury đã lấy thông tin được tạo ra cho một mục đích và chuyển đổi nó thành một cái gì đó khác nữa.

Phương pháp của ông, gần tương tự với các kỹ thuật dữ-liệu-lớn ngày hôm nay, thật đáng kinh ngạc nếu xét rằng nó đã được thực hiện chỉ với giấy và bút chì. Câu chuyện của ông làm nổi bật mức độ của việc sử dụng dữ liệu trước thời đại số hóa. Ngày nay chúng ta có xu hướng kết hợp hai thứ này, nhưng điều quan trọng là giữ chúng tách biệt. Để có được một sự hình dung đầy đủ hơn về cách dữ liệu được trích xuất từ những nơi ít ngờ đến nhất, hãy xem một ví dụ hiện đại hơn.

Đánh giá tư thế của con người là môn nghệ thuật cả khoa học của Shigeomi Koshimizu, một giáo sư tại Học viện cao cấp Nhật Bản về Công nghệ ở Tokyo. Ít ai nghĩ rằng cách một người ngồi lại chứa đựng thông tin, nhưng thật ra là có. Khi một người đang ngồi, những yếu tố như đường nét của cơ thể, tư thế, và phân

phối trọng lượng... đều có thể được định lượng và lập bảng. Koshimizu và đội ngũ kỹ sư của ông chuyển đổi các phần phía sau cơ thể thành dữ liệu bằng cách đo áp lực tại 360 điểm khác nhau từ cảm biến trong ghế ngồi xe và lập chỉ số mỗi điểm trên thang điểm từ 0 đến 256. Kết quả là mỗi cá nhân sẽ có một mã số duy nhất. Trong một thử nghiệm, hệ thống đã có thể phân biệt giữa khá nhiều người với độ chính xác 98 phần trăm.

Nghiên cứu kể trên không phải thứ ngớ ngẩn. Công nghệ này đang được phát triển thành một hệ thống chống trộm trong xe hơi. Một chiếc xe được trang bị công nghệ này sẽ nhận ra một người nào đó, khác với người lái xe đã được xác nhận, đang ngồi sau tay lái. Khi đó nó sẽ yêu cầu một mật khẩu để cho phép tiếp tục lái xe hoặc ngắt động cơ. Việc chuyển các tư thế ngồi thành dữ liệu đã tạo ra một dịch vụ khả thi và một công việc kinh doanh có khả năng sinh lợi. Và tính hữu dụng của nó có thể đi xa hơn cả việc ngăn chặn hành vi trộm cắp xe hơi. Ví dụ các dữ liệu tổng hợp có thể tiết lộ những manh mối về sự liên hệ giữa tư thế của người lái và mức an toàn giao thông, chẳng hạn như tư thế ngồi trước khi xảy ra tai nạn. Hệ thống cũng có thể cảm nhận được khi người lái xe có dấu hiệu mệt mỏi để gửi một cảnh báo hoặc tự động nhấn phanh. Và có thể nó không chỉ ngăn chặn một vụ ăn cắp xe mà còn xác định được kẻ trộm từ cặp mông của hắn (có thể nói như vậy).

Giáo sư Koshimizu đã chọn một thứ chưa bao giờ được xem như dữ liệu - hoặc thậm chí từng được hình dung rằng có khả năng cung cấp thông tin - và chuyển đổi nó thành một dạng số liệu. Tương tự như vậy, thuyền trưởng Maury đã chọn những tài liệu dường như rất ít có khả năng sử dụng để trích thông tin, biến nó thành dữ liệu vô cùng hữu ích. Việc này giúp các thông tin được sử dụng theo cách mới mẻ và tạo ra một giá trị độc đáo.

Từ “dữ liệu” mang nghĩa “đã có” trong tiếng Latin, theo nét nghĩa là một “điều thực tế”. Nó đã trở thành tiêu đề của một công

trình kinh điển của Euclid, trong đó ông giải thích hình học từ những gì được biết đến hoặc có thể được chứng minh là được biết đến. Ngày nay dữ liệu ám chỉ một cái gì đó cho phép nó được ghi lại, phân tích, và tổ chức. Chưa có thuật ngữ chính xác cho các loại chuyển đổi như của thuyền trưởng Maury và giáo sư Koshimizu. Vì vậy, hãy tạm gọi chúng là dữ liệu hóa (datafication). Dữ liệu hóa một hiện tượng là đặt nó trong một dạng định lượng để nó có thể được phân tích và lập bảng.

Một lần nữa, điều này rất khác với việc số hóa - quá trình chuyển đổi thông tin dạng tương tự thành những số 0 và 1 của mã nhị phân để máy tính có thể xử lý được, số hóa không phải là thứ đầu tiên chúng ta làm với máy tính. Thời kỳ ban đầu của cuộc cách mạng máy tính là tính toán, như từ nguyên của nó cho thấy. Chúng ta sử dụng máy để làm các phép tính toán từng đòi hỏi rất nhiều thời gian nếu bằng các phương pháp trước đây: chẳng hạn như bảng quỹ đạo tên lửa, tổng điều tra dân số, và dự báo thời tiết. Chỉ sau đó mới đến việc lấy nội dung tương tự và số hóa nó. Do đó khi Nicholas Negroponte của MIT Media Lab xuất bản cuốn sách mang tính bước ngoặt của ông năm 1995 tên là *BeingDigital,* một trong những chủ đề lớn của ông là sự chuyển đổi từ các nguyên tử sang các bit. Về căn bản, chúng ta đã số hóa văn bản trong những năm 1990. Gần đây hơn, khi khả năng lưu trữ, sức mạnh xử lý, và băng thông đã tăng lên, chúng ta đã thực hiện nó với các hình dạng nội dung khác như hình ảnh, video, và âm nhạc.

Ngày nay có một niềm tin tuyệt đối trong các chuyên gia công nghệ rằng dữ liệu lớn bắt nguồn từ cuộc cách mạng Silicon. Nhưng tất nhiên không phải vậy. Các hệ thống công nghệ thông tin hiện đại chắc chắn đã làm cho dữ liệu lớn trở nên khả thi, nhưng cốt lõi của việc chuyển đổi sang dữ liệu lớn là sự tiếp nối của cuộc tìm kiếm cổ xưa của loài người để đo lường, ghi lại và phân tích thế giới. Cuộc cách mạng IT là điều hiển nhiên khắp xung quanh chúng ta, nhưng sự nhấn mạnh chủ yếu vẫn trên

chữ *T (technology)*, công nghệ. Đã tới lúc phải thay đổi cách nhìn của chúng ta để tập trung vào chữ *I (information)*, thông tin. Để nắm bắt thông tin có thể định lượng, để dữ liệu hóa, chúng ta cần biết cách đo lường và ghi lại những gì chúng ta đo. Điều này đòi hỏi các công cụ thích hợp. Nó cũng đòi hỏi một khao khát được định lượng và ghi chép lại. Cả hai đều là điều kiện tiên quyết của việc dữ liệu hóa, và chúng ta đã phát triển các yếu tố cơ sở cần thiết cho dữ liệu hóa từ nhiều thế kỷ trước buổi bình minh của thời đại kỹ thuật số.

Định lượng thế giới

Khả năng ghi thông tin là một trong những đường ranh phân giới giữa xã hội nguyên thủy và xã hội tiên tiến. Đếm và đo lường cơ bản về chiều dài và trọng lượng là một trong những công cụ mang tính khái niệm lâu đời nhất của các nền văn minh sớm.

Vào thiên niên kỷ thứ ba trước Công nguyên, ý tưởng về ghi chép lại thông tin đã tiến bộ đáng kể trong vùng thung lũng Indus, Ai Cập và Lưỡng Hà. Độ chính xác tăng lên, cũng như việc sử dụng đo lường trong cuộc sống hàng ngày. Sự phát triển của chữ viết ở vùng Lưỡng Hà đã mang đến một phương pháp chính xác cho việc theo dõi sản xuất và các giao dịch kinh doanh. Ngôn ngữ viết cho phép các nền văn minh sớm đo lường được những yếu tố thực tại, ghi lại chúng, và truy tìm chúng sau này. Kết hợp với nhau, việc đo lường và ghi nhận đã hỗ trợ việc tạo ra dữ liệu. Chúng là những nền tảng đầu tiên của dữ liệu hóa.

Điều này tạo ra khả năng tái tạo hoạt động của con người. Ví dụ các tòa nhà có thể được sao lại từ hồ sơ các kích thước và vật liệu của chúng. Nó cũng cho phép thử nghiệm: một kiến trúc sư hay một nhà xây dựng có thể thay đổi một số kích thước nhất định trong khi vẫn giữ những kích thước khác không thay đổi, tạo ra một thiết kế mới - mà sau đó có thể được ghi lại. Các giao dịch thương mại có thể được ghi nhận, vì vậy người ta biết sản lượng

từ một vụ thu hoạch hay trên một cánh đồng (và bao nhiêu bị nhà nước lấy đi trong các loại thuế). Định lượng cho phép dự đoán và do đó lập kế hoạch, ngay cả khi chỉ là thô như đơn giản đoán xem mùa thu hoạch năm tiếp theo có dồi dào như các năm trước không. Nó cho phép các đối tác trong một giao dịch ghi nhận những gì họ còn nợ nhau. Nếu không có đo lường và ghi chép thì có thể đã không có tiền, vì sẽ không có được dữ liệu để hỗ trợ nó.

Qua nhiều thế kỷ, việc đo lường được mở rộng từ chiều dài và trọng lượng đến diện tích, khối lượng và thời gian. Vào đầu thiên niên kỷ thứ nhất sau Công nguyên, các tính năng chính của đo lường đã có ở phương Tây. Nhưng có một thiếu sót đáng kể về cách thức đo lường của các nền văn minh sớm. Nó không được tối ưu hóa cho việc tính toán, thậm chí cả những phép tính toán tương đối đơn giản. Hệ thống đếm với các chữ số La Mã không phù hợp cho việc phân tích số. Nếu không có một hệ thống cơ số mười hay số thập phân, các phép nhân và chia những số lớn là rất khó khăn ngay cả đối với các chuyên gia, và các phép đơn giản cộng và trừ sẽ khó hiểu đối với hầu hết những người còn lại.

Một hệ thống số khác đã được phát triển ở Ấn Độ vào khoảng thế kỷ thứ nhất sau Công nguyên. Nó đã lan đến Ba Tư và được cải thiện, rồi sau đó được chuyển sang những người Ả Rập, là những người đã tinh chỉnh nó rất nhiều. Nó là cơ sở của các chữ số Ả Rập chúng ta sử dụng ngày nay. Cuộc Thập tự chinh có thể đã hủy diệt các vùng đất mà người châu Âu xâm chiếm, nhưng kiến thức đã di chuyển từ Đông sang Tây, và có lẽ sự di chuyển quan trọng nhất là chữ số Ả Rập. Giáo hoàng Sylvester II, người từng nghiên cứu chúng, đã ủng hộ việc sử dụng chúng vào cuối thiên niên kỷ thứ nhất. Tới thế kỷ XII, các văn bản tiếng Ả Rập mô tả hệ thống này đã được dịch sang tiếng Latin và lan khắp châu Âu. Kết quả là toán học đã cất cánh.

Ngay cả trước khi chữ số Ả Rập đến với châu Âu, việc tính toán

đã được cải thiện thông qua các bàn tính. Đó là những khay nhẵn, trên đó các thẻ được đặt để biểu thị số lượng. Bằng việc trượt các thẻ trong những vùng nhất định, người ta có thể cộng hoặc trừ. Tuy nhiên, phương pháp này có những hạn chế nghiêm trọng. Thật khó để tính toán những con số rất lớn và rất nhỏ cùng một lúc. Quan trọng nhất, những con số trên bàn tính này không rõ ràng. Một bước di chuyển sai hoặc một va chạm bất cẩn có thể thay đổi một con số, dẫn đến những kết quả không chính xác. Bàn tính có thể được chấp nhận cho việc tính toán, nhưng chúng rất kém để ghi chép. Và cách duy nhất để ghi lại, lưu trữ các số hiển thị trên các bàn tính là chuyển chúng trở lại vào chữ số La Mã không mấy hiệu quả. (Những người châu Âu chưa bao giờ được tiếp xúc với các bàn tính của phương Đông - trong nhận thức muộn màng thì đó là một điều tốt, vì các thiết bị này có thể đã kéo dài việc sử dụng chữ số La Mã ở phương Tây.)

Toán học đã mang lại cho dữ liệu một ý nghĩa mới - bây giờ nó có thể được *phân tích*, chứ không chỉ được ghi lại và trích xuất. Việc áp dụng rộng rãi chữ số Ả Rập ở châu Âu đã phải mất hàng trăm năm, từ khi chúng xuất hiện vào thế kỷ XII đến cuối thế kỷ XVI. Vào thời điểm đó, các nhà toán học tự hào rằng họ có thể tính toán sáu lần nhanh hơn bằng chữ số Ả Rập so với bàn tính. Những gì cuối cùng đã giúp làm cho chữ số Ả Rập thành công là sự tiến hóa của một công cụ khác của dữ liệu hóa: kế toán kép.

Các nhà kế toán đã phát minh ra sổ sách kế toán vào thiên niên kỷ thứ ba trước công nguyên. Trong khi sổ sách kế toán phát triển qua nhiều thế kỷ sau đó, chủ yếu nó vẫn là một hệ thống ghi chép một giao dịch cụ thể ở một nơi. Những gì nó không thể làm được là cho các nhà kế toán và các ông chủ của họ biết một cách dễ dàng vào bất cứ lúc nào những gì họ quan tâm nhất: liệu một tài khoản cụ thể hoặc toàn bộ một công việc làm ăn có lợi nhuận hay không. Điều này bắt đầu thay đổi vào thế kỷ XIV, khi các nhà kế toán tại Ý bắt đầu ghi các giao dịch sử dụng hai mục, một cho các khoản có và một cho các khoản nợ, do đó tổng thể

các tài khoản là cân bằng, vẻ đẹp của hệ thống này là nó cho phép dễ dàng nhìn thấy lợi nhuận và thua lỗ. Và đột nhiên dữ liệu vô tri vô giác bắt đầu biết nói.

Ngày nay kế toán kép thường chỉ được dùng nhờ công dụng của nó đối với kế toán và tài chính. Nhưng nó cũng đại diện cho một bước ngoặt trong sự phát triển của việc sử dụng dữ liệu. Nó cho phép thông tin được ghi lại theo hình thức các “hạng mục” liên kết các tài khoản với nhau. Nó vận hành bằng một bộ quy tắc về cách ghi dữ liệu như thế nào - một trong những ví dụ sớm nhất của việc ghi chuẩn của thông tin. Một kế toán viên có thể nhìn vào sổ sách của người khác và hiểu được chúng. Nó được tổ chức để thực hiện một loại hình cụ thể của việc truy vấn dữ liệu - tính toán lợi nhuận hoặc lỗ cho mỗi tài khoản - nhanh chóng và đơn giản. Và nó cung cấp những bằng chứng kiểm toán của các giao dịch để dữ liệu được dễ dàng theo dõi hơn. Các chuyên gia công nghệ có lẽ sẽ đánh giá cao nó hôm nay: nó có tính năng “sửa lỗi” được tích hợp. Nếu một bên của sổ kế toán trông không ổn, người ta có thể kiểm tra các mục tương ứng bên kia.

Tuy nhiên, cũng như chữ số Ả Rập, kế toán kép không phải là một thành công ngay lập tức. Hai trăm năm sau khi phương pháp này lần đầu tiên được nghĩ ra, nó đã cần một nhà toán học và một gia đình thương gia để làm thay đổi lịch sử của dữ liệu hóa.

Nhà toán học đó là một tu sĩ dòng Phanxicô, Luca Pacioli. Năm 1494 ông xuất bản một cuốn sách giáo khoa, viết cho đại chúng, về toán học và ứng dụng thương mại của nó. Cuốn sách này là một thành công lớn và thật ra có vai trò như cuốn sách giáo khoa toán học của thời đó. Nó cũng là cuốn sách đầu tiên sử dụng chữ số Ả Rập, và do đó sự phổ biến của nó đã tạo điều kiện cho việc chấp nhận chữ số Ả Rập ở châu Âu. Tuy nhiên, đóng góp lâu dài nhất của nó là phần dành cho sổ sách kế toán, trong đó Pacioli giải thích cặn kẽ hệ thống kế toán kép. Trong nhiều thập kỷ kế

tiếp, tư liệu về sổ sách kế toán đã được xuất bản riêng bằng sáu ngôn ngữ, và nó đã là tài liệu tham khảo tiêu chuẩn về chủ đề này trong nhiều thế kỷ.

Còn về gia đình thương gia, đó là những thương nhân Venetian nổi tiếng và những nhà bảo hộ nghệ thuật: gia tộc Medici. Trong thế kỷ XVI, họ đã trở thành những chủ ngân hàng có ảnh hưởng nhất ở châu Âu, một phần không nhỏ vì họ đã sử dụng một phương pháp ưu việt để ghi dữ liệu: hệ thống kép. Cùng với nhau, sách giáo khoa của Pacioli và sự thành công của Medici trong việc áp dụng nó đã chốt lại chiến thắng của kế toán kép - và rộng hơn đã thiết lập được việc sử dụng chữ số Ả Rập ở phương Tây.

Song song với những tiến bộ trong việc ghi chép dữ liệu, những cách thức đo lường thế giới - biểu thị thời gian, khoảng cách, diện tích, khối lượng, và trọng lượng - đã tiếp tục đạt được độ chính xác ngày càng tăng. Lòng khao khát muốn hiểu được bản chất của sự vật thông qua định lượng đã định hình khoa học trong thế kỷ XIX, khi các học giả phát minh ra các công cụ và các đơn vị mới mẻ để đo và ghi lại dòng điện, áp suất không khí, nhiệt độ, tần số âm thanh... Đó là một thời đại mà tuyệt nhiên tất cả mọi thứ đều phải được xác định, lập ranh giới, và ký hiệu. Niềm đam mê đó còn đi xa tới mức đo sọ người để đo lường cho khả năng trí tuệ của họ. May mắn là cái môn giả-khoa-học nghiên cứu về sọ đã hầu như chết yểu, nhưng mong muốn định lượng mọi thứ cứ ngày càng tăng.

Việc đo lường hiện thực và ghi dữ liệu được phát triển mạnh là do sự kết hợp của các công cụ và một tư duy luôn sẵn sàng tiếp thu. Sự kết hợp này chính là mảnh đất màu mỡ từ đó dữ liệu hóa hiện đại đã phát triển. Các thành tố cho dữ liệu hóa đã tồn tại, mặc dù trong một thế giới của dữ liệu dạng tương tự, nó vẫn còn đắt đỏ và tốn thời gian. Trong nhiều trường hợp nó đòi hỏi dường như sự kiên nhẫn vô hạn, hoặc ít nhất là một sự cống hiến

lâu dài, như việc quan sát các ngôi sao và các hành tinh về đêm đầy nhọc nhằn của Tycho Brahe trong những năm 1500.

Trong một số ít các trường hợp dữ liệu hóa thành công, như lược đồ hàng hải của trung tá Maury, nó thường là một sự trùng hợp may mắn: chẳng hạn Maury được giao một công việc bàn giấy nhưng với quyền truy cập vào một kho tàng nhật ký hàng hải. Tuy nhiên, bất cứ khi nào dữ liệu hóa thật sự thành công, nó đều tạo ra được những giá trị khổng lồ từ các thông tin cơ bản và mở ra những hiểu biết phi thường.

Sự xuất hiện của máy tính đã mang đến những thiết bị đo lường và lưu trữ kỹ thuật số giúp dữ liệu hóa trở nên hiệu quả hơn rất nhiều. Nó cũng giúp khám phá được những giá trị tiềm ẩn từ việc phân tích toán học đối với dữ liệu. Tóm lại, số hóa tăng tốc cho dữ liệu hóa. Nhưng nó không phải là một sự thay thế. Hoạt động số hóa - chuyển thông tin dạng tương tự thành dạng máy tính đọc được - tự nó không phải là dữ liệu hóa.

Khi từ ngữ trở thành dữ liệu

Sự khác biệt giữa số hóa và dữ liệu hóa trở nên rõ ràng khi chúng ta xem xét một lĩnh vực mà cả hai hiện tượng đã xảy ra và so sánh kết quả của chúng: sách. Năm 2004 Google đã công bố một kế hoạch táo bạo. Họ sẽ lấy tất cả các trang của tất cả các cuốn sách mà họ có được (trong khuôn khổ pháp luật về bản quyền) và cho phép tất cả mọi người trên toàn thế giới tìm kiếm và truy cập miễn phí qua Internet. Để đạt được điều này công ty hợp tác với một số thư viện lớn nhất và uy tín nhất trên thế giới và phát triển những máy quét có thể tự động lật các trang, để việc quét hàng triệu cuốn sách vừa có thể thực hiện được và vừa khả thi về mặt tài chính.

Đầu tiên, Google *số hóa* văn bản: từng trang được quét và ghi trong một tập tin hình ảnh có độ phân giải kỹ thuật số cao, được

lưu trữ trên máy chủ của Google. Trang sách được chuyển thành một bản sao kỹ thuật số có thể dễ dàng được bất kỳ ai ở bất kỳ đâu truy cập thông qua Web. Tuy nhiên, việc truy cập sẽ đòi hỏi người đọc phải biết cuốn sách nào có thông tin mình quan tâm, hoặc phải đọc nhiều để tìm ra thông tin cần thiết. Người ta không thể tìm kiếm văn bản theo từ khóa, hoặc phân tích nó, bởi vì văn bản chưa được dữ liệu hóa. Tất cả những gì Google có là những hình ảnh mà chỉ con người mới có thể biến đổi thành thông tin hữu ích - bằng cách đọc.

Dù nó vẫn là một công cụ tuyệt vời - một Thư viện Alexandria kỹ thuật số hiện đại, toàn diện hơn bất kỳ thư viện nào trong lịch sử - Google vẫn muốn nhiều hơn nữa. Họ hiểu rằng thông tin chứa đựng những giá trị mà chỉ có thể được chuyển tải một khi nó được dữ liệu hóa. Và do vậy Google đã sử dụng phần mềm nhận dạng ký tự quang học để đọc một hình ảnh kỹ thuật số và nhận dạng ra các chữ cái, từ, câu, và đoạn văn trên đó. Kết quả là văn bản đã được dữ liệu hóa chứ không chỉ là một hình ảnh kỹ thuật số của trang sách.

Bây giờ các thông tin trên trang sách mới có thể được sử dụng không chỉ cho người đọc, mà còn cho các máy tính để xử lý và cho các thuật toán để phân tích. Dữ liệu hóa làm cho văn bản có thể lập chỉ mục và do đó có thể tìm kiếm được. Và nó cho phép một dòng phân tích văn bản bất tận. Bây giờ chúng ta có thể khám phá khi nào thì những từ hoặc cụm từ nhất định được sử dụng lần đầu tiên, hoặc trở nên phổ biến. Đó chính là thứ kiến thức làm sáng tỏ sự lan truyền của những ý tưởng và quá trình tiến hóa của tư duy con người qua nhiều thế kỷ và trong nhiều ngôn ngữ khác nhau. Bạn có thể tự thử nghiệm. Ngram Viewer của Google (http://books.google.com/ngrams) sẽ tạo ra một đồ thị của việc sử dụng các từ hoặc cụm từ theo thời gian, bằng cách sử dụng toàn bộ chỉ mục Sách của Google như một nguồn dữ liệu. Trong vòng vài giây chúng ta khám phá ra rằng cho đến năm 1900 thuật ngữ “nhân quả” được sử dụng thường xuyên

hơn “tương quan”, nhưng sau đó tỷ lệ này đã đảo ngược. Chúng ta có thể so sánh phong cách văn bản và xác định được tác giả khi có tranh chấp tác quyền. Dữ liệu hóa cũng giúp cho việc phát hiện đạo văn trong các công trình hàn lâm trở nên dễ dàng hơn, kết quả là một số chính trị gia châu Âu, trong đó có một bộ trưởng quốc phòng Đức, đã bị buộc phải từ chức.

Ước tính có khoảng 130 triệu đầu sách đã được xuất bản kể từ khi in ấn được phát minh ra vào giữa thế kỷ XV. Đến năm 2012, bảy năm sau khi Google bắt đầu dự án sách, họ đã sao chụp hơn 20 triệu đầu sách, hơn 15 phần trăm di sản in ấn của thế giới - một khối lượng đáng kể. Điều này đã tạo ra một ngành học mới được gọi là “Culturomics”: từ vựng học tính toán để cố gắng hiểu hành vi con người và các xu hướng văn hóa thông qua việc phân tích định lượng các văn bản số hóa.

Trong một nghiên cứu, các chuyên gia tại Đại học Harvard khảo sát hàng triệu cuốn sách (tương đương với hơn 500 tỷ từ) và phát hiện ra rằng chỉ có chưa đến một nửa số lượng các từ tiếng Anh xuất hiện trên sách là có trong các từ điển. Thay vào đó, họ viết, sự dồi dào của từ ngữ “bao gồm cả từ vựng‘ngoài lề’ vốn không được ghi chép trong các nguồn tham khảo chuẩn”. Hơn nữa, bằng việc phân tích theo thuật toán các tài liệu tham khảo về nghệ sĩ Marc Chagall, người có các tác phẩm bị Đức Quốc xã cấm vì là người Do Thái, các nhà nghiên cứu đã chỉ ra rằng sự đàn áp hoặc kiểm duyệt một ý tưởng hoặc cá nhân để lại “dấu vết có thể định lượng được”. Từ ngữ cũng giống như hóa thạch được bọc trong các trang viết thay vì trầm tích đá. Các nhà nghiên cứu culturomics có thể khai thác chúng như các nhà khảo cổ.

Việc chuyển từ ngữ thành dữ liệu mở ra rất nhiều công dụng. Tất nhiên, dữ liệu có thể được con người sử dụng để đọc, còn máy móc dùng chúng để phân tích. Nhưng là mẫu mực của một công ty dữ-liệu-lớn, Google biết rằng thông tin còn có nhiều khả năng tiềm ẩn khác, có thể giúp ích cho bộ sưu tập của mình và cho dữ

liệu hóa. Vì vậy, Google khéo léo sử dụng các văn bản được dữ liệu hóa từ dự án quét sách để cải thiện dịch vụ dịch máy của mình. Như đã giải thích trong Chương Ba, hệ thống sẽ lấy những cuốn sách được dịch và phân tích những từ và cụm từ nào được các dịch giả sử dụng như những lựa chọn thay thế từ một ngôn ngữ sang một ngôn ngữ khác. Hiểu biết được điều này thì sau đó có thể xử lý việc dịch như một vấn đề toán học khổng lồ, với các máy tính tìm ra xác suất để xác định từ nào là thay thế tốt nhất cho từ kia giữa các ngôn ngữ.

Tất nhiên Google không phải là tổ chức duy nhất mơ ước mang đến sự phong phú của di sản in ấn của thế giới vào thời đại máy tính, và nó hầu như không phải là nơi đầu tiên thử việc này. Dự án Gutenberg, một sáng kiến tình nguyện để đưa các tác phẩm thuộc sở hữu công cộng lên trực tuyến sớm có từ năm 1971, nhằm giúp mọi độc giả dễ tiếp cận các văn bản này. Tuy nhiên, dự án đã không xem xét một chức năng phụ trợ của từ ngữ nên không xem chúng như dữ liệu. Tương tự như vậy, các nhà xuất bản trong nhiều năm qua đã thử nghiệm với các phiên bản sách điện tử. Họ cũng nhìn thấy giá trị cốt lõi của sách là nội dung, chứ không phải là dữ liệu - mô hình kinh doanh của họ dựa vào điều này. Vì vậy, họ không bao giờ sử dụng hoặc cho phép người khác sử dụng các dữ liệu vốn có trong văn bản của một cuốn sách. Họ không bao giờ thấy sự cần thiết, hoặc đánh giá cao tiềm năng đó.

Nhiều công ty hiện nay đang cạnh tranh để chiếm lĩnh thị trường sách điện tử. Amazon, với máy đọc sách điện tử Kindle của mình, dường như là người dẫn đầu sớm. Nhưng đây là một lĩnh vực mà chiến lược của Amazon và Google khác nhau rất nhiều. Amazon đã dữ liệu hóa sách - nhưng không giống như Google, họ đã thất bại trong việc khai thác những chức năng mới của văn bản với vai trò dữ liệu. Jeff Bezos, người sáng lập và giám đốc điều hành của công ty, đã thuyết phục hàng trăm nhà xuất bản để phát hành sách của họ dưới dạng Kindle. Sách Kindle

không phải được tạo từ ảnh của trang sách. Nếu như vậy, người đọc sẽ không thể thay đổi kích thước chữ hoặc hiển thị trang sách cả trên màn hình màu và trắng đen. Văn bản được dữ liệu hóa, không chỉ là số hóa. Thật ra, Amazon đã làm việc đó cho hàng triệu cuốn sách mới, những gì Google đang cố gắng cẩn thận đạt được đối với nhiều cuốn sách cũ hơn.

Tuy nhiên, khác với dịch vụ tuyệt vời của Amazon với "những từ ngữ quan trọng về mặt thống kê" - trong đó sử dụng các thuật toán để tìm liên kết giữa các chủ đề của sách mà bình thường có thể không rõ ràng - nhà bán lẻ trực tuyến đã không tận dụng sự giàu có của từ ngữ cho phân tích dữ-liệu-lớn. Amazon xem việc kinh doanh sách của mình là dựa trên nội dung độc giả xem, chứ không phải trên phân tích văn bản dữ liệu hóa. Và để công bằng, Amazon có thể phải đối mặt với những hạn chế từ các nhà xuất bản bảo thủ về việc Amazon có thể sử dụng thông tin chứa đựng trong các cuốn sách của họ như thế nào. Google, một cậu bé dữ-liệu-lớn hư hỏng sẵn sàng đẩy xa các giới hạn, không nhận thấy những hạn chế như vậy: bánh mì của Google được phết bơ bằng những cú nhấp chuột, chứ không phải bằng việc truy cập các đầu sách của người sử dụng. Có lẽ là công bằng khi nói rằng ít nhất trong lúc này, Amazon hiểu được giá trị của việc số hóa nội dung, trong khi Google hiểu được giá trị của việc dữ liệu hóa nó.

Khi vị trí trở thành dữ liệu

Một trong những phần cơ bản nhất của thông tin trong thế giới này chính là... bản thân thế giới. Nhưng qua gần hết lịch sử, lĩnh vực không gian chưa bao giờ được định lượng hoặc sử dụng ở dạng dữ liệu. Vị trí địa lý của thiên nhiên, các vật thể, và con người tất nhiên cấu thành thông tin. Dãy núi là ở đó; người là ở đây. Nhưng để trở nên hữu ích nhất, thông tin này phải được trở thành dữ liệu. Việc dữ liệu hóa vị trí đòi hỏi một vài điều kiện tiên quyết. Chúng ta cần một phương pháp để đo mỗi inch vuông của bề mặt Trái đất. Chúng ta cần một cách chuẩn hóa để ghi chú

các phép đo. Chúng ta cần một công cụ để theo dõi và ghi lại các dữ liệu. Định lượng, tiêu chuẩn hóa, thu thập. Chỉ khi đó chúng ta mới có thể lưu trữ và phân tích vị trí không chỉ như nơi chốn, mà như dữ liệu.

Ở phương Tây, việc định lượng vị trí bắt đầu với người Hy Lạp. Khoảng năm 200 trước Công nguyên, Eratosthenes đã phát minh ra một hệ thống các đường lưới để phân ranh giới vị trí, giống như vĩ độ và kinh độ. Nhưng cũng giống như rất nhiều ý tưởng hay từ thời cổ đại, việc thực hành đã phai nhạt dần theo thời gian. Một thiên niên kỷ rưỡi sau, khoảng năm 1400 sau Công nguyên, một bản sao *Geographia* của Ptolemy đến Florence từ Constantinople, cũng giống như thời kỳ Phục hưng và buôn bán vận chuyển đã khơi dậy mối quan tâm đến khoa học và bí quyết từ người xưa. Luận thuyết của Ptolemy đã gây một sự náo động, và những bài học cũ của ông đã được áp dụng để giải quyết những thách thức trong hàng hải hiện đại. Từ đó, bản đồ xuất hiện với kinh độ, vĩ độ và tỷ lệ. Hệ thống sau đó đã được một nhà bản đồ học người Flanders, Gerardus Mercator, cải thiện vào năm 1570, cho phép các thủy thủ lập một tuyến đường thẳng trong một thế giới hình cầu.

Mặc dù thời điểm đó đã có phương tiện để ghi lại vị trí, nhưng chưa có định dạng được chấp nhận phổ biến để chia sẻ những thông tin này. Một hệ thống nhận diện chung là cần thiết, cũng giống như Internet hưởng lợi từ tên miền để làm những thứ như email hoạt động được một cách phổ dụng. Việc tiêu chuẩn hóa kinh độ và vĩ độ mất một thời gian dài. Cuối cùng nó được ghi nhận vào năm 1884 tại Hội nghị quốc tế Meridian ở Washington, DC, nơi mà 25 quốc gia đã chọn Greenwich, Anh, như kinh tuyến chính và điểm không của kinh độ (người Pháp, vốn tự xem mình là những nhà lãnh đạo về các tiêu chuẩn quốc tế, bỏ phiếu trắng). Trong những năm 1940 hệ tọa độ Universal Transverse Mercator (UTM) đã được tạo ra, phân chia thế giới thành 60 vùng để tăng độ chính xác.

Vị trí không gian địa lý bây giờ có thể được xác định, ghi nhận, đo đếm, phân tích, và chuyển tải trong một định dạng số chuẩn. Vị trí có thể được dữ liệu hóa. Nhưng vì chi phí để đo và ghi lại các thông tin trong môi trường dữ liệu ở dạng tương tự sẽ cao, nên nó hiếm khi được thực hiện. Để việc dữ liệu hóa diễn ra, người ta phải phát minh các công cụ đo vị trí với giá rẻ. Cho đến những năm 1970, cách duy nhất để xác định vị trí địa lý là sử dụng các điểm mốc, các chòm sao thiên văn, hoặc công nghệ radio định vị hạn chế.

Một sự thay đổi lớn đã xảy ra vào năm 1978, khi vệ tinh đầu tiên trong số 24 vệ tinh tạo nên hệ thống định vị toàn cầu (GPS) được phóng lên. Các thiết bị thu trên mặt đất có thể lập lưới tam giác vị trí của chúng bằng cách ghi nhận sự khác biệt về thời gian cần để nhận được một tín hiệu từ các vệ tinh cách xa 12.600 dặm trên không. Được Bộ Quốc phòng Hoa Kỳ phát triển, hệ thống lần đầu tiên được mở ra cho các mục đích phi quân sự trong những năm 1980 và được vận hành đầy đủ vào những năm 1990. Độ chính xác của nó được tăng cường cho các ứng dụng thương mại một thập kỷ sau đó. Chính xác đến từng mét, GPS đánh dấu thời điểm một phương thức đo vị trí, giấc mơ của các nhà hàng hải, các nhà làm bản đồ, và các nhà toán học từ thời cổ đại, cuối cùng đã được hợp nhất với các phương tiện kỹ thuật để thành công một cách nhanh chóng, với giá (tương đối) rẻ, và không yêu cầu bất kỳ kiến thức chuyên môn nào.

Tuy nhiên, các thông tin phải thực sự được tạo ra. Không có gì ngăn Eratosthenes và Mercator ước tính vị trí của họ mỗi phút trong ngày, nếu họ thích. Dù khả thi nhưng điều đó lại phi thực tế. Tương tự như vậy, những máy thu GPS ban đầu vừa phức tạp vừa đắt, thích hợp cho một chiếc tàu ngầm nhưng không phải cho tất cả mọi người ở mọi thời điểm. Tuy nhiên điều này đã thay đổi, nhờ vào sự phổ biến của các chip rẻ tiền nhúng trong các tiện ích kỹ thuật số. Giá của một mô-đun GPS giảm từ hàng trăm đôla trong những năm 1990 xuống khoảng một đôla ngày

nay với số lượng lớn. Thường chỉ mất vài giây để GPS xác định được một vị trí, và tọa độ được chuẩn hóa. Vì vậy, 37° 14’ 06” Bắc, 115° 48’ 40” Tây chỉ có thể nghĩa là ta đang ở một căn cứ quân sự siêu bí mật của Mỹ ở một vùng hẻo lánh của bang Nevada được gọi là “Vùng 51”, nơi người ngoài hành tinh (có lẽ!) đang bị giam giữ.

Ngày nay GPS chỉ là một trong số nhiều hệ thống để nắm bắt vị trí. Các hệ thống vệ tinh đối thủ đang được tiến hành tại Trung Quốc và châu Âu. Và thậm chí độ chính xác tốt hơn có thể được thiết lập bởi lập lưới tam giác giữa các tháp di động hoặc các bộ định tuyến wifi để xác định vị trí dựa trên cường độ tín hiệu, vì GPS không hoạt động bên trong nhà hoặc giữa các tòa nhà cao tầng. Điều đó giúp giải thích tại sao các công ty như Google, Apple và Microsoft: đã thiết lập những hệ thống vị trí địa lý riêng của họ để bổ sung cho GPS. Các xe Street View của Google thu thập thông tin bộ định tuyến wifi khi họ chụp ảnh, và iPhone là một “spyPhone” (điện thoại do thám) thu thập dữ liệu vị trí và wifi và gửi nó trở lại Apple, mà người dùng không hề nhận ra. (Điện thoại Android của Google và hệ điều hành di động của Microsoft cũng thu thập loại dữ liệu này.)

Không chỉ người mà các vật thể cũng có thể bị theo dõi. Với những module vô tuyến đặt bên trong xe, việc dữ liệu hóa vị trí sẽ làm thay đổi các ý tưởng về bảo hiểm. Dữ liệu cho biết một cách chi tiết về thời gian, địa điểm, và khoảng cách xe chạy thực tế để định giá rủi ro tốt hơn. Ở Mỹ và Anh, người lái xe có thể mua bảo hiểm xe định giá theo thực tế xe được lái ở đâu và lúc nào, chứ không chỉ trả giá hàng năm theo tuổi tác, giới tính và hồ sơ quá khứ. Cách tiếp cận này để định giá bảo hiểm tạo ra những ưu đãi cho hành vi tốt. Nó thay đổi bản chất của bảo hiểm từ dựa trên sự tổng hợp nguy cơ sang một cái gì đó dựa trên hành động cá nhân. Việc theo dõi cá nhân thông qua chiếc xe cũng thay đổi bản chất của các chi phí cố định, như đường giao thông và cơ sở hạ tầng khác, bằng cách gắn việc sử dụng những tài

nguyên này với những người lái xe và những người khác "tiêu thụ" chúng. Người ta đã không thể làm điều này trước khi chuyển vị trí địa lý trở thành một dạng dữ liệu liên tục cho tất cả mọi người và tất cả mọi thứ - nhưng đó là thế giới chúng ta đang đi tới.

Ví dụ UPS sử dụng dữ liệu 'Vị-trí-địa-lý" theo nhiều cách. Xe của hãng được trang bị cảm biến, mô-đun vô tuyến, và GPS để trụ sở có thể dự đoán sự cố động cơ, như chúng ta đã thấy trong chương trước. Hơn nữa, nó cho phép công ty biết nơi chốn của xe tải trong trường hợp chậm trễ, để giám sát nhân viên, và theo dõi hành trình của họ để tối ưu hóa các tuyến đường.



Phim minh họa cơ chế phân tích của UPS

Chương trình phân tích này có tác động rất đặc biệt. Theo Jack

Levis, giám đốc quản lý quy trình của UPS, năm 2011 UPS đã thu ngắn các tuyến đường cho xe của công ty tới 30 triệu dặm, tiết kiệm 3 triệu gallon nhiên liệu và 30 ngàn tấn carbon dioxide khí thải. Nó cũng cải thiện tính an toàn và hiệu quả: thuật toán tạo ra các tuyến đường với ít đoạn rẽ qua các giao lộ, yếu tố vốn thường dẫn đến tai nạn, lãng phí thời gian, và tiêu thụ nhiều nhiên liệu hơn vì xe thường xuyên phải dừng trước khi rẽ.

“Việc dự báo đã cho chúng ta kiến thức”, Levis của hãng UPS nói. “Nhưng phía sau kiến thức là một cái gì đó nhiều hơn nữa: sự khôn ngoan và sáng suốt. Tại một thời điểm nào đó, hệ thống sẽ thông minh tới mức nó sẽ dự đoán các vấn đề và sửa chữa chúng trước khi người dùng nhận ra rằng có điều gì đó sai”.

Đáng chú ý nhất là việc dữ liệu hóa vị trí theo thời gian được áp dụng cho con người. Trong nhiều năm qua, các nhà khai thác vô tuyến đã thu thập và phân tích thông tin để nâng tầm dịch vụ của mạng lưới của họ. Nhưng dữ liệu ngày càng được sử dụng nhiều cho các mục đích khác và được thu thập bởi bên thứ ba cho những dịch vụ mới. Ví dụ một số ứng dụng điện thoại thông minh thu thập thông tin vị trí cho dù bản thân ứng dụng có một tính năng dựa trên địa điểm hay không. Trong những trường hợp khác, ứng dụng chỉ được dùng để xây dựng một doanh nghiệp tận dụng kiến thức về địa điểm của người sử dụng. Một ví dụ là Foursquare, cho phép mọi người “đăng nhập” tại các địa điểm yêu thích của họ. Nó kiếm được thu nhập từ các chương trình khách hàng trung thành, giới thiệu nhà hàng, và các dịch vụ khác liên quan đến vị trí.

Khả năng thu thập dữ liệu vị trí địa lý của người sử dụng đang trở nên vô cùng giá trị. Ở mức độ cá nhân, nó giúp cho việc quảng cáo nhắm đến mục tiêu dựa trên việc khách hàng đang ở đâu và dự đoán sẽ đi tới đâu. Hơn nữa, thông tin có thể được tổng hợp để cho biết các xu hướng. Ví dụ việc tích lũy dữ liệu vị trí cho phép các công ty phát hiện ùn tắc giao thông mà không cần

trông thấy những chiếc xe, nhờ số lượng và tốc độ của các máy điện thoại di chuyển trên một đường cao tốc tiết lộ thông tin này. Công ty AirSage xử lý 15 tỷ bản ghi thông tin vị trí địa lý mỗi ngày từ sự di chuyển của hàng triệu thuê bao điện thoại di động để tạo các báo cáo giao thông thời gian thực ở hơn 100 thành phố trên khắp nước Mỹ. Hai công ty vị trí địa lý khác, Sense Networks và Skyhook, có thể sử dụng dữ liệu vị trí để cho biết các khu vực của một thành phố có cuộc sống về đêm nhộn nhịp nhất, hoặc để ước tính có bao nhiêu người đã có mặt tại một cuộc biểu tình.

Tuy nhiên, những ứng dụng phi thương mại của vị trí địa lý mới chứng tỏ tầm quan trọng nhất. Sandy Pentland, Giám đốc Phòng thí nghiệm Động lực học Con người của MIT, và Nathan Eagle đã cùng nhau đi tiên phong trong lĩnh vực họ gọi là “khai thác thực tế”. Nó đề cập đến việc xử lý những lượng lớn dữ liệu từ điện thoại di động để đưa ra những kết luận và dự đoán về hành vi con người. Trong một nghiên cứu, việc phân tích các chuyển động và các mô hình cuộc gọi đã cho phép họ xác định thành công những người đã mắc bệnh cúm trước khi bản thân họ biết rằng họ bị bệnh. Trong trường hợp của một dịch cúm chết người, khả năng này có thể cứu hàng triệu sinh mạng bằng cách cho các nhân viên y tế công biết các khu vực bị ảnh hưởng nhất vào bất cứ lúc nào. Nhưng nếu đặt vào những bàn tay vô trách nhiệm thì sức mạnh của “khai thác thực tế” có thể gây nên những hậu quả khủng khiếp, như chúng ta sẽ thấy sau này.

Eagle, người sáng lập của công ty khởi động dữ liệu vô tuyến Jana, đã tập hợp dữ liệu điện thoại di động từ hơn 200 nhà khai thác trong hơn 100 quốc gia - khoảng 3,5 tỷ người ở châu Mỹ Latin, châu Phi, và châu Âu - để trả lời những câu hỏi mà các nhà quản lý tiếp thị quan tâm, như mỗi tuần một hộ gia đình giặt bao nhiêu lần. Nhưng ông cũng sử dụng dữ liệu lớn để kiểm tra các câu hỏi như các thành phố phát triển thịnh vượng như thế nào. Ông và một đồng nghiệp đã kết hợp dữ liệu vị trí trên các thuê

bao điện thoại di động trả trước ở châu Phi với số tiền họ bỏ ra khi họ có nhiều tiền nhất trong tài khoản. Giá trị này tương quan mạnh với thu nhập: người giàu hơn mua nhiều phút hơn tại một thời điểm. Nhưng một trong những phát hiện ngược lại với lẽ thường mà Eagle thu được là những khu nhà ổ chuột, không chỉ là những khu trung tâm của sự nghèo nàn, mà còn hoạt động như những bàn đạp kinh tế. Điều quan trọng là những ứng dụng gián tiếp của dữ liệu vị trí không có gì liên quan tới việc định tuyến của truyền thông di động, mục đích ban đầu mà vì nó thông tin đã được tạo ra. Thay vào đó, khi vị trí được dữ liệu hóa, những công dụng mới sẽ nảy mầm và giá trị mới có thể được tạo ra.

Khi việc tương tác trở thành dữ liệu

Biên giới tiếp theo của dữ liệu hóa sẽ mang tính cá nhân hơn: các mối quan hệ, kinh nghiệm, và tâm trạng của chúng ta. Ý tưởng của dữ liệu hóa là xương sống của nhiều công ty truyền thông xã hội trên Web. Các diễn đàn mạng xã hội không chỉ đơn giản cung cấp cho chúng ta một cách để tìm và giữ liên lạc với bạn bè và đồng nghiệp, chúng lấy các yếu tố vô hình trong cuộc sống hàng ngày của chúng ta và biến thành dữ liệu có thể được sử dụng để làm những điều mới mẻ. Facebook dữ liệu hóa các mối quan hệ. Chúng luôn luôn tồn tại và cấu thành thông tin, nhưng chưa bao giờ được chính thức định nghĩa như là dữ liệu cho đến khi có "đồ thị xã hội" của Facebook. Twitter giúp dữ liệu hóa cảm xúc bằng cách tạo ra một cách dễ dàng cho người dùng ghi lại và chia sẻ những điều bận tâm của họ, mà trước đó đã bị "cuốn trôi" vào những con gió của thời gian. Linkedln dữ liệu hóa các kinh nghiệm chuyên môn trong quá khứ của chúng ta (giống như Maury đã chuyển những cuốn nhật ký hàng hải cũ), biến thông tin đó thành những dự đoán về hiện tại và tương lai: người mà chúng ta có thể biết, hoặc một công việc mà chúng ta có thể mong muốn.

Những cách sử dụng dữ liệu như vậy vẫn ở dạng phôi thai. Trong trường hợp của Facebook, điều này đã được thực hiện kiên nhẫn một cách khôn ngoan, vì công ty hiểu rằng việc tiết lộ quá nhiều mục đích mới cho dữ liệu của người sử dụng quá sớm có thể sẽ làm họ hoảng sợ. Bên cạnh đó, Facebook vẫn đang điều chỉnh mô hình kinh doanh của mình (và chính sách bảo mật) cho số lượng và loại hình thu thập dữ liệu nó muốn tiến hành. Do đó đa phần những lời chỉ trích mà nó phải đối mặt tập trung vào những thông tin nào nó có khả năng thu thập hơn là về những gì nó đã thực sự làm được với dữ liệu đó. Facebook có khoảng hơn một tỷ người sử dụng vào năm 2013, những người đã kết nối với nhau thông qua hơn 100 tỷ mối quan hệ bạn bè. Kết quả là đồ thị xã hội thu được đại diện cho hơn 10 phần trăm tổng dân số thế giới, được dữ liệu hóa và dễ tiếp cận đối với duy nhất một công ty.

Các ứng dụng tiềm năng của nó rất có triển vọng. Một số công ty mới thành lập đã cân nhắc việc tùy biến các đồ thị xã hội để sử dụng như những chỉ báo cho việc thiết lập điểm số tín dụng. Nó xuất phát từ ý tưởng là những con chim cùng loại thường tụ đàn: người thận trọng kết bạn với những người thận trọng, trong khi những kẻ trác táng thì lòng thòng với nhau. Nếu mở rộng, Facebook có thể là FICO tiếp theo, một cơ quan lập điểm tín dụng. Các bộ dữ liệu phong phú từ các công ty truyền thông xã hội cũng có thể tạo nên cơ sở của các doanh nghiệp mới, vượt xa việc chia sẻ hình ảnh, cập nhật trạng thái, và “thích”.

Twitter cũng nhận thấy dữ liệu của mình được sử dụng theo nhiều cách thú vị. Với một số người, việc 400 triệu tweet ngắn gọn được gửi đi mỗi ngày trong năm 2012 bởi hơn 140 triệu người sử dụng hàng tháng có vẻ ít nhiều giống như sự ba hoa rỗng tuếch ngẫu nhiên. Và, trên thực tế, chúng thường chỉ là như vậy. Tuy nhiên, công ty này tạo điều kiện cho việc dữ liệu hóa những suy nghĩ, tâm trạng, và mối tương tác của mọi người, những thứ chưa hề được thu lượm trước đó. Twitter đã thỏa thuận với hai công ty, Data-Sift và Gnip, để bán quyền truy cập

vào dữ liệu. (Mặc dù tất cả các tweet là tài sản công cộng, việc truy cập vào "suối nguồn" phải tốn chi phí.) Nhiều doanh nghiệp phân tích cú pháp các tweet, đôi khi sử dụng một kỹ thuật gọi là phân tích cảm xúc, để thu thập toàn bộ phản hồi của khách hàng hoặc đánh giá tác động của chiến dịch tiếp thị.

Hai quỹ phòng hộ, Derwent Capital ở London và MarketPsych ở California, đã bắt đầu phân tích các văn bản được dữ liệu hóa của tweet như các tín hiệu cho đầu tư vào thị trường chứng khoán. (Các chiến lược kinh doanh thực tế của họ được giữ bí mật. Thay vì đổ tiền vào các công ty được quảng cáo rùm beng, có lẽ họ đã đầu tư cho sự suy thoái của chúng.) Cả hai công ty bây giờ bán các thông tin cho các nhà đầu tư. MarketPsych hợp tác với Thomson Reuters để cung cấp không dưới 18.864 chỉ số riêng biệt trên 119 quốc gia, được cập nhật từng phút, dựa trên các trạng thái cảm xúc như lạc quan, u ám, vui vẻ, sợ hãi, giận dữ, và ngay cả các chủ đề như đổi mới, kiện tụng, và xung đột.

Dữ liệu được sử dụng bởi con người không nhiều như bởi máy tính: các thần đồng toán học của Wall Street, được gọi là "những cây sào", cắm dữ liệu vào các mô hình thuật toán của họ để tìm kiếm các mối tương quan vô hình có thể tận dụng để tạo ra lợi nhuận. Tần số của tweet về một chủ đề có thể dự đoán những điều khác nhau, chẳng hạn như doanh thu phòng vé của Hollywood, theo một trong những cha đẻ của phân tích mạng xã hội, Bernardo Huberman. Ông và một đồng nghiệp ở HP đã phát triển một mô hình xem xét tốc độ các tweet mới được đăng. Với điều này, họ đã có thể dự báo về thành công của một bộ phim tốt hơn so với các dự báo quen thuộc khác.

Nhưng còn có thể làm được nhiều thứ hơn thế nữa. Các tin nhắn Twitter bị giới hạn trong 140 ký tự, nhưng các siêu dữ liệu - tức "thông tin về thông tin" - kết hợp với mỗi tweet lại phong phú. Nó bao gồm 33 mục riêng biệt. Một số mục dường như không hữu ích, như "hình nền" trên trang Twitter của người sử dụng

hoặc phần mềm họ dùng để truy cập vào dịch vụ. Nhưng những siêu dữ liệu khác lại vô cùng thú vị, chẳng hạn như ngôn ngữ của người sử dụng, vị trí địa lý của họ, số lượng và tên của những người họ "theo dõi", hoặc những người "theo dõi" họ. Một nghiên cứu được đăng trên tạp chí *Science* năm 2011, phân tích 509 triệu tweet qua hai năm từ 2,4 triệu người ở 84 quốc gia, cho thấy tâm trạng của họ tuân theo các khuôn mẫu theo ngày và theo tuần tương tự nhau dù ở các nền văn hóa khác nhau trên thế giới - một điều không thể phát hiện được trước đây. Tâm trạng đã được dữ liệu hóa.

Việc dữ liệu hóa không chỉ liên quan đến việc biểu thị thái độ và tình cảm thành một hình thức có thể phân tích được, mà cả hành vi của con người. Điều này khó theo dõi được theo cách khác, đặc biệt là trong bối cảnh của cộng đồng rộng lớn hơn và các nhóm con bên trong nó. Nhà sinh vật học Marcel Salathé của Đại học Penn State cùng kỹ sư phần mềm Shashank Khandelwal đã phân tích các tweet và phát hiện ra rằng thái độ của nhiều người về tiêm chủng cũng phù hợp với khả năng họ đã thực sự chích ngừa cúm. Tuy nhiên, điều quan trọng là nghiên cứu của họ sử dụng siêu dữ liệu về ai đã kết nối với ai trong số những người "theo dõi" nhau trên Twitter để đi thêm một bước xa hơn. Họ nhận thấy rằng những phân nhóm người chưa chích ngừa có thể vẫn tồn tại. Điều làm cho nghiên cứu này trở nên đặc biệt là trong khi các nghiên cứu khác, chẳng hạn như Xu hướng Dịch cúm của Google, sử dụng dữ liệu tổng hợp để đánh giá tình trạng sức khỏe của các cá nhân, thì phân tích cảm xúc của Salathé đã thực sự dự đoán *hành vi* liên quan đến sức khỏe.

Những phát hiện sớm trên cho thấy dữ liệu hóa chắc chắn sẽ đi tiếp tới đâu. Cũng giống như Google, các mạng truyền thông xã hội như Facebook, Twitter, Linkedln, Foursquare, và nhiều mạng khác đang ngồi trên một rương khổng lồ các thông tin được dữ liệu hóa, mà một khi được phân tích, sẽ rọi ánh sáng lên các động lực xã hội ở tất cả mọi cấp độ, từ các cá nhân đến toàn bộ xã hội.

Dữ liệu hóa tất cả mọi thứ

Chỉ cần vận dụng chút trí tưởng tượng, ta có thể hình dung một kho tàng đủ mọi thứ có thể được chuyển thành dạng dữ liệu - và khiến chúng ta kinh ngạc. Với cùng một tinh thần như công trình của giáo sư Koshimizu về dáng điệu, IBM đã được cấp bằng sáng chế ở Mỹ vào năm 2012 về "Bảo đảm an toàn nhà cửa bằng công nghệ máy tính dựa trên bề mặt". Đó là bằng sáng chế cho một sàn nhà cảm ứng, phần nào giống như một màn hình điện thoại thông minh khổng lồ. Triển vọng của việc sử dụng nó rất khả quan. Sàn nhà kiểu này có thể xác định các vật thể trên đó. Về cơ bản, nó có thể biết bật đèn một phòng hoặc mở cửa khi có người đi vào. Tuy nhiên, quan trọng hơn, nó có thể xác định các cá nhân theo trọng lượng của họ hay cách họ đứng và đi. Nó có thể biết nếu một người nào đó ngã và không đứng dậy được, một tính năng quan trọng cho người cao tuổi. Các nhà bán lẻ có thể biết được dòng di chuyển của khách mua trong các cửa hàng của họ. Một khi sàn nhà được dữ liệu hóa thì chẳng có "nóc nhà" nào giới hạn được các ứng dụng tiềm tàng của nó.

Việc dữ liệu hóa càng nhiều càng tốt không phải là chuyện xa vời như ta tưởng. Chẳng hạn số lượng "những-người-tự-theo-dõi-mình" là nhỏ tại thời điểm hiện nay nhưng sẽ ngày càng tăng. Nhờ điện thoại thông minh và công nghệ điện toán giá rẻ, việc dữ liệu hóa các hành vi quan trọng nhất của cuộc sống chưa bao giờ dễ dàng hơn. Rất nhiều công ty mới thành lập đã giúp mọi người theo dõi giấc ngủ của họ bằng cách đo sóng não suốt đêm. Công ty Zeo đã tạo ra cơ sở dữ liệu lớn nhất thế giới về giấc ngủ và những khác biệt về số giai đoạn "ngủ động mắt nhanh" (REM) của cả nam giới và nữ giới. Asthmapolis đã gắn một cảm biến lên một ống hít cho bệnh nhân hen suyễn để theo dõi vị trí thông qua GPS, tập hợp thông tin giúp công ty nhận rõ những yếu tố từ môi trường gây nên cơn hen suyễn, chẳng hạn như cự ly tới một số loại cây trồng nhất định.

Các công ty Fitbit và Jawbone giúp mọi người đo hoạt động thể chất và giấc ngủ của họ. Một công ty khác, Basis, cho phép người mang vòng đeo tay theo dõi các dấu hiệu sống của họ, trong đó có nhịp tim và độ dẫn của da - những thông số đo được sự căng thẳng. Việc có được dữ liệu ngày càng trở nên dễ dàng hơn và đơn giản hơn bao giờ hết. Năm 2009 Apple đã được cấp bằng sáng chế cho việc thu thập dữ liệu về mức ôxy trong máu, nhịp tim và nhiệt độ cơ thể bằng tai nghe của nó.

Có rất nhiều thứ để học hỏi từ việc dữ liệu hóa cách thức cơ thể một con người hoạt động. Các nhà nghiên cứu tại Đại học Gjovik ở Na Uy và Derawi Biometrics đã phát triển một ứng dụng cho điện thoại thông minh có thể phân tích dáng đi của một cá nhân trong khi đi bộ và sử dụng thông tin này như một hệ thống bảo mật để mở khóa điện thoại. Trong khi đó hai giáo sư tại Viện Nghiên cứu Công nghệ Georgia, Robert Delano và Brian Parise, đang phát triển một ứng dụng điện thoại thông minh được gọi là iTrem sử dụng đồng hồ gia tốc gắn trong điện thoại để theo dõi các chấn động cơ thể cho bệnh Parkinson và những rối loạn thần kinh khác, ứng dụng này là một lợi ích cho cả bác sĩ và bệnh nhân. Nó cho phép bệnh nhân bỏ qua những cuộc kiểm tra tốn kém tại phòng khám, nó cũng cho phép các chuyên gia y tế giám sát từ xa tình trạng của bệnh nhân và phản ứng của họ với các bước điều trị. Theo các nhà nghiên cứu ở Kyoto, một điện thoại thông minh chỉ kém hiệu quả chút ít khi đo các chấn động so với đồng hồ gia tốc ba trục sử dụng trong ngành y tế, vì vậy người ta có thể yên tâm sử dụng nó. Một lần nữa, một chút hỗn độn đã chiến thắng tính chính xác.

Trong hầu hết các trường hợp, chúng ta nắm bắt thông tin và chuyển thành dạng dữ liệu để cho phép nó được tái sử dụng. Điều này có thể xảy ra gần như ở khắp mọi nơi và gần như đối với tất cả mọi thứ. GreenGoose, một công ty mới thành lập ở San Francisco, bán các cảm biến nhỏ xíu phát hiện chuyển động, có thể được đặt trên các vật thể để theo dõi xem chúng được sử

dụng nhiều bao nhiêu. Nếu đặt cảm biến trên một hộp chỉ nha khoa, một bình tưới nước, hoặc một cái chuồng mèo thì có thể dữ liệu hóa được việc vệ sinh răng miệng, chăm sóc cây trồng hoặc vật nuôi. Người ta hăng hái với những gì liên quan đến Internet một phần là vì chuyện lập mạng lưới, nhưng cũng còn vì việc dữ liệu hóa tất cả những gì xung quanh chúng ta.

Khi thế giới đã được dữ liệu hóa, tiềm năng sử dụng thông tin về cơ bản chỉ bị giới hạn bởi sự sáng tạo của mỗi người. Maury đã dữ liệu hóa những chuyên đi trước đây của thủy thủ thông qua việc lập bảng bằng tay rất siêng năng, và do đó đã mở khóa cho những hiểu biết và giá trị phi thường. Ngày nay chúng ta có các công cụ (số liệu thống kê và các thuật toán) và thiết bị cần thiết (những bộ xử lý kỹ thuật số và bộ nhớ) để thực hiện những công việc tương tự nhanh hơn, với quy mô lớn, và trong nhiều bối cảnh khác nhau. Trong thời đại của dữ liệu lớn, thậm chí những bộ phận xấu xí cũng có nhiều mặt tốt đẹp để sử dụng.

Chúng ta đang ở trung tâm của một dự án cơ sở hạ tầng tuyệt vời mà theo nghĩa nào đó là đối thủ của những dự án trong quá khứ, từ cống dẫn nước La Mã tới Bách khoa toàn thư của sự Khai sáng. Chúng ta không đánh giá hết điều này bởi vì dự án ngày nay là rất mới mẻ, bởi vì chúng ta đang ở ngay giữa nó, và bởi vì không giống như nước chảy trong cống, sản phẩm lao động của chúng ta là vô hình. Dự án đó là dữ liệu hóa. Giống như những tiến bộ cơ sở hạ tầng khác, nó sẽ mang lại những thay đổi cơ bản cho xã hội. cống dẫn nước đã tạo điều kiện cho các thành phố phát triển; in ấn đã tạo điều kiện cho Khai sáng; và báo chí đã thúc đẩy sự phát triển của nhà nước độc lập. Nhưng những cơ sở hạ tầng này tập trung vào các dòng chảy - của nước, của kiến thức. Điện thoại và Internet cũng vậy. Ngược lại, dữ liệu hóa đại diện cho một sự làm giàu quan trọng đối với hiểu biết của con người.

Với sự trợ giúp của dữ liệu lớn, chúng ta sẽ không còn xem thế giới như một chuỗi các diễn biến được giải thích như những hiện

tượng tự nhiên hoặc xã hội, mà như một vũ trụ bao gồm chủ yếu là thông tin. Trong hơn một thế kỷ, các nhà vật lý đã đề nghị như vậy - rằng không phải các nguyên tử mà thông tin mới là cơ sở của tất cả mọi thứ. Phải thừa nhận rằng điều này có vẻ bí hiểm. Tuy nhiên, thông qua dữ liệu hóa, trong nhiều trường hợp chúng ta có thể nắm bắt và tính toán các khía cạnh vật chất và phi vật thể của sự sống và tác động lên chúng, trên một quy mô toàn diện hơn nhiều.

Việc xem thế giới như thông tin, như đại dương dữ liệu có thể được khám phá với bề rộng và chiều sâu lớn nhất từ trước đến nay, cho chúng ta một cái nhìn về thực tế mà chúng ta chưa hề có. Đây là một quan điểm có thể thâm nhập tất cả các lĩnh vực của đời sống. Ngày nay, chúng ta là một xã hội định lượng bởi chúng ta cho rằng có thể hiểu được thế giới bằng những con số và toán học. Và chúng ta thừa nhận kiến thức có thể được truyền tải qua thời gian và không gian vì ý tưởng của chữ viết ăn rất sâu vào trí não. Trong tương lai, có lẽ các thế hệ tiếp theo sẽ có một "ý thức dữ-liệu-lớn". Khái niệm về chuyển đổi vô số chiều kích của thực tế thành dữ liệu có thể dường như mới mẻ đối với hầu hết mọi người hiện nay. Nhưng trong tương lai, chúng ta chắc chắn sẽ xem nó như một sự hiển nhiên (điều thú vị là nó trở lại nguồn gốc sâu xa của thuật ngữ "dữ liệu").

Theo thời gian, tầm vóc ý nghĩa của dữ liệu hóa có thể khiến sự phát minh ra cống dẫn nước và báo chí trở thành nhỏ nhoi. Nó có thể sánh ngang với in ấn và Internet, khi mang đến cho chúng ta những phương tiện để sắp xếp lại thế giới theo một cách định lượng và có thể phân tích được. Tuy nhiên, tại thời điểm này, những người tiến bộ nhất trong dữ liệu hóa lại đang thuộc giới kinh doanh, nơi dữ liệu lớn đang được sử dụng để tạo ra các hình thức giá trị mới. Đây cũng chính là chủ đề của chương kế tiếp.

6. GIÁ TRỊ

VÀO CUỐI NHỮNG NĂM 1990, Web đã nhanh chóng trở thành một nơi chốn phóng túng, khó chịu và kém thân thiện. "Thư rác" tràn ngập các hộp thư điện tử và các diễn đàn trực tuyến. Năm 2000, Luis von Ahn, một thanh niên 22 tuổi, vừa tốt nghiệp đại học, đã có một ý tưởng để giải quyết vấn đề: bắt buộc những ai đăng ký phải chứng minh họ là con người. Do vậy, anh tìm cái gì đó rất dễ dàng để con người làm nhưng lại rất khó khăn cho máy.

Anh đã đưa ra ý tưởng hiển thị những chữ nguệch ngoạc, khó đọc trong quá trình đăng ký. Con người sẽ có thể đọc được chúng và gõ vào chính xác trong một vài giây, nhưng máy móc sẽ bối rối. Yahoo áp dụng phương pháp của anh và giảm được mối họa của thư rác ngay lập tức. Von Ahn gọi sáng tạo của mình là Captcha (viết tắt của Completely Automated Public Turing Test to Tell Computers and Humans Apart - Phép kiểm tra Turing hoàn toàn tự động để phân biệt máy tính với con người). Năm năm sau, hàng triệu Captcha đã được gõ vào mỗi ngày.

Captcha đã mang lại cho von Ahn sự nổi tiếng và công việc giảng dạy về khoa học máy tính tại Đại học Carnegie Mellon sau khi anh có bằng tiến sỹ. Nó cũng đóng vai trò giúp anh, khi mới 27 tuổi, nhận được một trong những giải thưởng uy tín cho "thiên tài" của Quỹ MacArthur với nửa triệu đôla. Tuy nhiên khi nhận ra mình chịu trách nhiệm cho việc hàng triệu người lãng phí rất nhiều thời gian mỗi ngày để gõ vào những chữ nguệch ngoạc gây phiền nhiễu - nhưng sau đó chẳng được dùng để làm gì - anh thấy như vậy chẳng thông minh cho lắm.

Tìm cách để đưa toàn bộ sức mạnh tính toán của con người vào sử dụng hiệu quả hơn, von Ahn đã đưa ra một phiên bản kế nhiệm thích hợp có tên ReCaptcha. Thay vì gõ vào các chữ cái ngẫu nhiên, người ta gõ vào hai từ, thuộc trong số các dự án quét văn bản mà chương trình nhận dạng ký tự quang học của máy

tính không thể hiểu được. Một từ được dùng để xác nhận điều những người dùng khác đã gõ vào và do đó là tín hiệu cho biết đó là một con người, còn từ kia là một từ mới cần làm rõ nghĩa. Để đảm bảo tính chính xác, hệ thống hiển thị cùng một từ không rõ nghĩa cho khoảng năm người khác nhau để họ gõ vào một cách chính xác trước khi hệ thống tin tưởng đó là đúng. Dữ liệu này có một ứng dụng chính - để chứng minh người dùng là con người - nhưng nó cũng có một mục đích thứ hai: để giải mã những chữ không rõ ràng trong các văn bản số hóa.

Giá trị mang lại là vô cùng lớn, khi ta nghĩ đến chi phí để thuê người thay thế. Mất khoảng 10 giây mỗi lần sử dụng, 200 triệu ReCaptcha mỗi ngày - mức hiện tại - sẽ nhân với nửa triệu giờ một ngày. Mức lương tối thiểu tại Hoa Kỳ là $7,25 một giờ vào năm 2012. Nếu dùng sức người để làm rõ nghĩa những từ mà máy tính không hiểu được, sẽ tốn 4 triệu đôla một ngày, hay hơn 1 tỷ đôla mỗi năm. Thay vào đó, von Ahn thiết kế một hệ thống để làm điều đó, và thật ra là miễn phí. Điều này có giá trị tới mức Google đã mua lại công nghệ từ von Ahn vào năm 2009, và sau đó cung cấp miễn phí cho bất kỳ trang web nào sử dụng. Ngày nay nó được đưa vào khoảng 200.000 trang web, trong đó có Facebook, Twitter, và Craigslist.



Phim minh họa ReCaptcha

Câu chuyện của ReCaptcha nhấn mạnh tầm quan trọng của việc tái sử dụng dữ liệu. Với dữ liệu lớn, giá trị của dữ liệu đang thay đổi. Giá trị của dữ liệu chuyển từ ứng dụng cơ bản sang các ứng dụng tiềm năng của nó. Điều này có những hệ quả sâu sắc. Nó ảnh hưởng đến cách các doanh nghiệp đánh giá dữ liệu họ nắm giữ và cho phép những ai truy cập. Nó cho phép, và có thể buộc các công ty phải thay đổi các mô hình kinh doanh của họ. Nó làm thay đổi cách thức các tổ chức suy nghĩ về dữ liệu và việc sử dụng nó.

Thông tin luôn luôn cần thiết cho các giao dịch thị trường. Ví dụ dữ liệu cho phép phát hiện giá cả, và đó là một tín hiệu để biết phải sản xuất bao nhiêu. Chúng ta hiểu rõ khía cạnh này của dữ liệu. Có một số loại thông tin từ lâu đã được giao dịch trên thị trường, ví dụ nội dung có trong các cuốn sách, bài viết, nhạc, và phim, hoặc thông tin tài chính như giá cổ phiếu. Những thứ này

đã được kết hợp với dữ liệu cá nhân trong vài thập kỷ qua. Những nhà môi giới chuyên ngành dữ liệu ở Hoa Kỳ như Acxiom, Experian và Equifax tính phí khá hào phóng đối với các hồ sơ đầy đủ của thông tin cá nhân về hàng trăm hàng triệu khách hàng. Nhờ Facebook, Twitter, Linkedln, và các nền tảng truyền thông xã hội khác, các kết nối cá nhân, ý kiến, sở thích, và mô hình cuộc sống hàng ngày của chúng ta đã tham gia vào vốn chung của thông tin cá nhân về chúng ta.

Một cách ngắn gọn, mặc dù dữ liệu từ lâu đã có giá trị, nó chỉ được xem như phụ trợ cho các hoạt động cốt lõi của một doanh nghiệp, hoặc bị giới hạn trong các phạm trù tương đối hẹp như sở hữu trí tuệ hoặc thông tin cá nhân. Ngược lại, trong thời đại của dữ liệu lớn, *tất cả* dữ liệu sẽ được xem là có giá trị, cả về nội dung và chính bản thân dữ liệu đó.

Khi nói "tất cả dữ liệu", chúng ta ám chỉ ngay cả thứ thô nhất, dường như hầu hết các bit trần trụi của thông tin. Hãy nghĩ tới các số đo từ một cảm biến nhiệt trên một máy ở công xưởng. Hoặc dòng thời gian thực của các tọa độ GPS, các số đo từ đồng hồ gia tốc, và các mức nhiên liệu từ một chiếc xe giao hàng - hay một đội xe gồm 60.000 chiếc. Hoặc hãy nghĩ tới hàng tỷ truy vấn tìm kiếm cũ, hoặc giá của từng ghế trên mỗi chuyến bay thương mại ở Hoa Kỳ trong nhiều năm qua.

Cho đến gần đây, không có cách dễ dàng để thu thập, lưu trữ, và phân tích những dữ liệu như vậy. Điều này hạn chế nghiêm trọng các cơ hội để tận dung giá trị tiềm năng của nó. Trong ví dụ nổi tiếng của Adam Smith về nhà sản xuất ghim, ông đã thảo luận về phân công lao động trong thế kỷ XVIII, phải đòi hỏi những người quan sát theo dõi tất cả các công nhân, không chỉ cho một nghiên cứu cụ thể, mà cho mọi thời điểm của mỗi ngày, lấy các số đo chi tiết, và đếm sản phẩm trên giấy dày với bút lông. Khi các nhà kinh tế cổ điển xem xét các yếu tố của sản xuất (đất đai, lao động và vốn), ý tưởng về khai thác dữ liệu hầu như

vắng bóng. Mặc dù chi phí để thu thập và sử dụng dữ liệu đã giảm trong hơn hai thế kỷ qua, cho đến khá gần đây nó vẫn còn tương đối tốn kém.

Điều làm cho thời đại của chúng ta khác biệt là rất nhiều hạn chế cố hữu về thu thập dữ liệu không còn nữa. Công nghệ đã đạt tới điểm mà những lượng lớn thông tin thường xuyên có thể được ghi nhận với giá rẻ. Dữ liệu có thể thường xuyên được thu thập một cách thụ động mà không cần nhiều nỗ lực hoặc thậm chí những đối tượng được ghi lại cũng không hề hay biết. Và bởi chi phí lưu trữ đã giảm rất nhiều, việc giữ lại dữ liệu thay vì loại bỏ nó trở nên dễ dàng hơn. Tất cả những thứ đó làm cho dữ liệu dễ tiếp cận và với chi phí thấp chưa từng có. Trong nửa thế kỷ qua, cứ hai năm thì chi phí lưu trữ kỹ thuật số lại giảm khoảng một nửa, trong khi mật độ lưu trữ đã tăng 50 triệu lần. Theo quan điểm của các công ty thông tin như Farecast hoặc Google - nơi các chất liệu thô đi vào ở một đầu của dây chuyền kỹ thuật số và thông tin đã được xử lý đi ra ở đầu kia - dữ liệu bắt đầu trông giống như một nguồn nguyên liệu mới của sản xuất.

Giá trị tức thời của hầu hết dữ liệu là hiển nhiên đối với những người thu thập. Thật ra, có lẽ họ tập hợp nó với một mục đích cụ thể. Các cửa hàng thu thập dữ liệu bán hàng để làm kế toán tài chính cho đúng. Các nhà máy theo dõi sản phẩm để đảm bảo chúng phù hợp với các tiêu chuẩn chất lượng. Các trang web ghi lại từng cú nhấp chuột của người dùng - đôi khi cả nơi con trỏ di chuyển - để phân tích và tối ưu hóa nội dung các trang web trình bày cho người ghé thăm. Những ứng dụng chính này của dữ liệu biện minh cho việc thu thập và xử lý nó. Khi lưu lại không chỉ những cuốn sách khách hàng mua mà cả các trang web họ đơn thuần nhìn vào, Amazon biết rằng họ sẽ sử dụng dữ liệu này để đưa ra những khuyên nghị cá nhân hóa. Tương tự như vậy, Facebook theo dõi việc “cập nhật trạng thái” và nhấn nút “like” của người dùng nhằm xác định những quảng cáo phù hợp nhất để hiển thị trên trang web của mình và kiếm tiền từ đó.

Không giống như những thứ vật chất - ví dụ thực phẩm chúng ta ăn, một cây nến cháy - giá trị của dữ liệu không giảm đi khi nó được sử dụng. Nó có thể được xử lý lại và xử lý lại nữa. Thông tin là thứ các nhà kinh tế gọi là hàng hóa “không-cạnh-tranh”: việc sử dụng của một người không cản trở việc sử dụng của người khác. Và thông tin không hao mòn khi sử dụng như các loại vật chất khác. Do đó Amazon có thể sử dụng dữ liệu từ các giao dịch quá khứ khi đưa ra những khuyến nghị cho khách hàng của mình - và sử dụng nó nhiều lần, không chỉ cho khách hàng đã tạo ra dữ liệu mà còn cho cả nhiều người khác nữa. Dữ liệu có thể được sử dụng nhiều lần cho cùng một mục đích. Quan trọng hơn, nó còn có thể được khai thác cho nhiều mục đích khác nhau. Điểm này rất quan trọng khi chúng ta cố gắng hiểu thông tin sẽ có giá trị bao nhiêu đối với chúng ta trong thời đại của dữ liệu lớn. Chúng ta thấy một số tiềm năng này đã trở thành hiện thực, như khi Walmart tìm kiếm cơ sở dữ liệu các hóa đơn bán hàng cũ và phát hiện ra mối tương quan hấp dẫn giữa các cơn bão và việc bán Pop-Tarts.

Tất cả những điều này cho thấy giá trị đầy đủ của dữ liệu là lớn hơn nhiều so với giá trị được trích xuất từ nó cho mục đích sử dụng ban đầu. Nó cũng có nghĩa là các công ty có thể khai thác dữ liệu một cách hiệu quả ngay cả khi việc sử dụng lần đầu hoặc mỗi lần tiếp theo chỉ mang lại một lượng nhỏ của giá trị, miễn là họ sử dụng dữ liệu nhiều lần.

“Giá trị tùy chọn” của dữ liệu

Để hiểu được ý nghĩa của việc tái sử dụng dữ liệu đối với giá trị cuối cùng của nó, hãy lấy ví dụ các xe hơi chạy điện. Khả năng để chúng thành công và trở thành một phương thức vận tải phụ thuộc vào một vô số các yếu tố hậu cần, mà tất cả đều liên quan tới hoạt động của bình điện. Người lái phải nạp được bình điện cho xe của họ một cách nhanh chóng và thuận tiện, và các công ty năng lượng cần đảm bảo rằng năng lượng dùng bởi những

chiếc xe này không làm mất ổn định lưới điện. Ngày nay, chúng ta có mạng phân phối khá hiệu quả các trạm xăng, nhưng chúng ta chưa hiểu được nhu cầu nạp điện và vị trí của các trạm cho xe hơi điện là như thế nào.

Điều đáng lưu tâm là vấn đề này không phải thiên về cơ sở hạ tầng mà thiên về thông tin. Và dữ liệu lớn là một phần quan trọng của giải pháp. Trong một thử nghiệm vào năm 2012, IBM đã làm việc với Công ty điện lực và khí Thái Bình Dương ở California và nhà sản xuất xe hơi Honda để thu thập một lượng lớn thông tin nhằm trả lời các câu hỏi cơ bản về thời gian và địa điểm xe điện sẽ nạp điện, và điều này có nghĩa gì đối với việc cung cấp năng lượng. IBM đã phát triển một mô hình dự đoán được xây dựng dựa trên rất nhiều yếu tố: lượng điện trong bình, vị trí của xe, thời gian trong ngày, và các chỗ đỗ có sẵn tại các trạm nạp điện gần đó. Nó kết hợp dữ liệu với mức tiêu thụ hiện tại từ lưới điện cũng như mô hình sử dụng năng lượng trong quá khứ. Việc phân tích các dòng lớn dữ liệu theo thời gian hiện tại và quá khứ từ nhiều nguồn cho phép IBM xác định những thời gian và địa điểm tối ưu cho người lái nạp bình điện xe của họ. Nó cũng tiết lộ nơi tốt nhất để xây dựng các trạm nạp. Cuối cùng, hệ thống sẽ phải tính đến chênh lệch giá tại các trạm nạp gần đó. Ngay cả dự báo thời tiết cũng được xem là một yếu tố: chẳng hạn trường hợp trời nắng và một trạm năng lượng mặt trời gần đó đầy ắp điện, nhưng dự báo thời tiết cho biết sắp có một tuần mưa nên các tấm pin mặt trời sẽ không vận hành.

Hệ thống lấy thông tin được tạo ra cho một mục đích và tái sử dụng nó cho một mục đích khác - nói cách khác, dữ liệu chuyển từ ứng dụng chính sang ứng dụng phụ. Điều này làm tăng giá trị của nó theo thời gian. Chỉ báo lượng điện của xe sẽ cho người lái biết khi nào thì cần nạp điện. Dữ liệu về sử dụng lưới điện được công ty dịch vụ tiện ích thu thập để quản lý sự ổn định của lưới điện. Đó là những ứng dụng chính. Cả hai bộ dữ liệu đều có những ứng dụng phụ - và giá trị mới - khi chúng được dùng cho

một mục đích hoàn toàn khác: xác định nên nạp điện khi nào và ở đâu, và nơi để xây dựng các trạm dịch vụ xe hơi điện. Thêm nữa, các thông tin phụ trợ được kết hợp, chẳng hạn như vị trí của xe và việc tiêu thụ lưới điện trong quá khứ. Và IBM xử lý dữ liệu không chỉ một lần mà còn xử lý lại và lại nữa, vì nó liên tục cập nhật hồ sơ tiêu thụ năng lượng của xe điện và ảnh hưởng của nó lên lưới điện.

Giá trị thực sự của dữ liệu giống như một tảng băng trôi nổi trên đại dương. Chỉ một phần nhỏ của nó là có thể được nhìn thấy ngay từ cái nhìn đầu tiên, trong khi phần lớn của nó bị ẩn bên dưới bề mặt. Các công ty sáng tạo hiểu được điều này có thể tận dụng được những giá trị và gặt hái những lợi ích tiềm năng rất lớn. Tóm lại, giá trị của dữ liệu phải được xem xét trên tất cả các khía cạnh nó có thể được sử dụng trong tương lai, chứ không chỉ đơn giản trong hiện tại. Chúng ta từng thấy điều này trong nhiều ví dụ đã được nhấn mạnh. Farecast khai thác dữ liệu từ vé máy bay bán trước đó để dự đoán giá vé tương lai. Google tái sử dụng các từ khóa tìm kiếm để khám phá sự lây lan của bệnh cúm. Maury đã sử dụng lại các nhật ký đi biển cũ để phát hiện những dòng hải lưu.

Tuy nhiên, tầm quan trọng của việc tái sử dụng dữ liệu vẫn chưa được đánh giá đầy đủ trong kinh doanh và xã hội. Rất ít nhà điều hành tại Con Edison ở New York có thể tưởng tượng được rằng thông tin về các cáp cũ hàng thế kỷ và các hồ sơ bảo trì có thể được sử dụng để ngăn ngừa tai nạn trong tương lai. Phải cần một thế hệ mới các nhà thống kê, và một làn sóng mới các phương pháp và công cụ để mở được khóa giá trị của dữ liệu. Ngay cả nhiều công ty Internet và công nghệ đến gần đây vẫn không hề biết việc tái sử dụng dữ liệu có thể có giá trị như thế nào.

Việc hình dung dữ liệu theo cách các nhà vật lý xem xét năng lượng cũng là một cách hay. Họ đề cập đến năng lượng “lưu trữ” hoặc “tiềm ẩn” tồn tại bên trong một đối tượng nhưng nằm im.

Hãy hình dung một lò xo bị nén hoặc một quả bóng dừng tại đỉnh của một ngọn đồi. Năng lượng trong các đối tượng này vẫn còn âm ỉ - tiềm ẩn - cho đến khi nó được giải phóng, chẳng hạn, khi lò xo được bung ra hoặc quả bóng được đẩy nhẹ để nó lăn xuống dốc. Lúc này năng lượng của các đối tượng đã trở thành "động" vì chúng đang chuyển động và tác dụng lên các đối tượng trong thế giới. Sau ứng dụng chính của nó, giá trị của dữ liệu vẫn còn tồn tại, nhưng nằm im, giống như lò xo hoặc quả bóng, cho đến khi dữ liệu được dùng cho một ứng dụng phụ và sức mạnh của nó lại được giải phóng. Trong thời đại dữ-liệu-lớn, cuối cùng chúng ta đã có được cách suy nghĩ, sự khéo léo, và các công cụ để khai thác giá trị tiềm ẩn của dữ liệu.

Cuối cùng, giá trị của dữ liệu là những gì người ta có thể đạt được từ tất cả các cách sử dụng nó. Những ứng dụng tiềm năng dường như vô hạn này cũng giống như những lựa chọn - không theo ý nghĩa của các công cụ tài chính, nhưng theo ý nghĩa thiết thực của sự lựa chọn. Giá trị của dữ liệu là tổng của các lựa chọn này: "giá trị lựa chọn" của dữ liệu, có thể nói như vậy. Trong quá khứ, một khi ứng dụng chính của dữ liệu đã đạt được, chúng ta thường nghĩ rằng dữ liệu đã hoàn thành mục đích của mình, và chúng ta sẵn sàng xóa nó, để cho nó mất đi. Xét cho cùng, dường như giá trị quan trọng đã được tận dụng. Trong thời đại dữ-liệu-lớn, dữ liệu giống như một mỏ kim cương huyền diệu vẫn tiếp tục sản xuất thêm lâu nữa sau khi giá trị chính của nó đã được khai thác. Có ba cách hiệu nghiệm để giải phóng giá trị tùy chọn của dữ liệu: tái sử dụng cơ bản, hợp nhất các tập dữ liệu, và tìm kiếm các "ích lợi kép".

TÁI SỬ DỤNG DỮ LIỆU

Một ví dụ điển hình của việc tái sử dụng sáng tạo dữ liệu là các từ khóa tìm kiếm. Thoạt đầu, thông tin có vẻ vô giá trị sau khi mục đích chính của nó đã được hoàn thành. Sự tương tác tạm thời giữa người sử dụng và công cụ tìm kiếm đưa ra một danh sách

các trang web và quảng cáo phục vụ một chức năng đặc biệt duy nhất cho thời điểm đó. Nhưng những truy vấn cũ có thể có giá trị bất thường. Hitwise, một công ty đo lường lưu lượng web thuộc sở hữu của nhà môi giới dữ liệu Experian, cho phép khách hàng khai thác lưu lượng tìm kiếm để tìm hiểu sở thích của người tiêu dùng. Các nhà tiếp thị có thể sử dụng Hitwise để hình dung liệu màu hồng sẽ lên ngôi trong mùa xuân này hay màu đen sẽ trở lại. Google đưa ra một phiên bản của bộ phân tích từ khóa tìm kiếm để mọi người kiểm tra. Nó đã khải động một dịch vụ dự báo kinh doanh với ngân hàng lớn thứ hai của Tây Ban Nha, BBVA, để xem xét ngành du lịch cũng như bán các chỉ số kinh tế thời gian thực dựa trên dữ liệu tìm kiếm. Ngân hàng Anh sử dụng các truy vấn tìm kiếm liên quan đến bất động sản để hình dung tốt hơn về việc giá nhà đất tăng hay giảm.

Các công ty thất bại trong việc đánh giá cao tầm quan trọng của tái sử dụng dữ liệu đã học được bài học của họ một cách khó khăn. Ví dụ, trong những ngày đầu của Amazon, họ đã ký một thỏa thuận với AOL để dùng công nghệ thương mại điện tử của AOL. Đối với hầu hết mọi người, nó trông giống như một thỏa thuận gia công bình thường. Nhưng những gì thực sự khiến Amazon quan tâm, như Andreas Weigend, cựu giám đốc khoa học của Amazon, giải thích là việc có được dữ liệu về những gì người dùng AOL đã xem và mua, điều sẽ cải thiện hiệu quả cho các khuyến nghị của Amazon. AOL tội nghiệp không hề nhận ra điều này. Họ chỉ nhìn thấy giá trị của dữ liệu trong mục đích sử dụng chính - bán hàng. Amazon thông minh biết họ có thể gặt hái lợi ích bằng cách đưa dữ liệu này vào một ứng dụng phụ.

Hoặc hãy xét trường hợp Google đã nhảy vào lĩnh vực nhận dạng giọng nói với GOOG-411 cho các danh sách tìm kiếm địa phương, thực hiện từ 2007 đến 2010. Người khổng lồ về tìm kiếm không có công nghệ nhận dạng giọng nói riêng của mình nên phải mua bản quyền. Google đạt được thỏa thuận với Nuance, công ty hàng đầu trong lĩnh vực này đã vui mừng gặp

được vị khách cao giá. Nhưng Nuance lúc đó là một gã ngốc về dữ-liệu-lớn: hợp đồng không chỉ định ai là người sẽ giữ các bản ghi dịch tiếng nói, và Google đã giữ chúng cho riêng mình. Việc phân tích dữ liệu cho phép người ta đánh giá xác suất để một đoạn số hóa nhất định của tiếng nói tương ứng với một từ cụ thể. Đây là điều quan trọng để cải thiện công nghệ nhận dạng giọng nói hoặc tạo ra một dịch vụ mới mẻ hoàn toàn. Thời điểm đó, Nuance cho rằng họ kinh doanh bản quyền phần mềm, chứ không phải phân tích dữ liệu. Ngay sau khi thấy lỗi của mình, họ mới bắt đầu có những thỏa thuận đáng chú ý với các nhà khai thác di động và các nhà sản xuất thiết bị cầm tay để sử dụng dịch vụ nhận dạng giọng nói của mình - để có thể thu thập được dữ liệu.

Giá trị trong việc tái sử dụng dữ liệu là tin tốt cho các tổ chức thu thập hoặc kiểm soát các bộ dữ liệu lớn nhưng hiện đang sử dụng chúng rất ít, chẳng hạn như những doanh nghiệp thường chủ yếu hoạt động ngoại tuyến (offline). Họ có thể ngồi trên những mỏ thông tin chưa được khai thác. Một số công ty có thể đã thu thập dữ liệu, sử dụng nó một lần (nếu có), và giữ nó ở đâu đó vi chi phí lưu trữ thấp - trong những "nấm mồ dữ liệu", như các nhà khoa học dữ liệu gọi những nơi thông tin cũ cư trú.

Các công ty Internet và công nghệ đang tiên phong khai thác hàng núi dữ liệu, vì họ thu thập được rất nhiều thông tin chỉ bằng cách hoạt động trực tuyến và đi trước các công ty khác trong việc phân tích nó. Nhưng tất cả các công ty đều được hưởng lợi. Các chuyên gia tư vấn tại McKinsey & Company cho biết một công ty hậu cần (giấu tên) nhận thấy trong quá trình cung cấp hàng hóa, nó đã tích lũy hàng đống thông tin về vận chuyển hàng hóa trên toàn cầu. Thấy được cơ hội, nó thành lập một bộ phận đặc biệt để bán dữ liệu tổng hợp ở dạng các dự báo kinh doanh và kinh tế. Nói cách khác, nó tạo ra một phiên bản ngoại tuyến của Google trong việc truy-vấn-tìm-kiếm-quá-khứ. Hoặc SWIFT, hệ thống liên ngân hàng toàn cầu để chuyển tiền,

đã phát hiện ra rằng các khoản thanh toán tương quan với các hoạt động kinh tế toàn cầu. Vì vậy, SWIFT cung cấp dự báo GDP dựa trên dữ liệu chuyển tiền đi qua mạng lưới của mình.

Một số doanh nghiệp, nhờ vào vị trí của họ trong chuỗi giá trị thông tin, có thể thu thập được những lượng lớn dữ liệu, mặc dù họ có ít nhu cầu ngay lập tức đối với dữ liệu hoặc không thành thạo trong việc sử dụng lại nó. Ví dụ các nhà khai thác điện thoại di động thu thập thông tin về địa điểm của các thuê bao để phân tuyến các cuộc gọi. Đối với những công ty này, dữ liệu như vậy chỉ có các mục đích kỹ thuật hạn hẹp. Nhưng nó có giá trị hơn khi được tái sử dụng bởi các công ty phân phối quảng cáo và chương trình khuyến mãi được cá nhân hóa dựa trên địa điểm. Đôi khi giá trị không đến từ các điểm dữ liệu riêng lẻ mà từ những gì chúng tiết lộ trong quá trình tổng hợp. Do đó các doanh nghiệp bán thông tin vị trí địa lý như AirSage và Sense Networks mà chúng ta đã thấy trong chương trước có thể bán thông tin về nơi mà người dân đang tụ tập vào một tối thứ Sáu hoặc nơi những chiếc xe đang phải bò chậm chạp trên đường. Những kiểu thông tin tổng hợp này có thể được sử dụng để xác định giá trị bất động sản hoặc giá bảng hiệu quảng cáo.

Ngay cả những thông tin tầm thường nhất cũng có thể có giá trị đặc biệt, nếu được áp dụng một cách đúng đắn. Hãy quay lại với các nhà khai thác điện thoại di động: họ lưu trữ về việc các điện thoại kết nối với các hạm cơ sở ở đâu và khi nào, với cường độ tín hiệu thế nào. Các nhà khai thác từ lâu đã sử dụng dữ liệu đó để tinh chỉnh hiệu suất mạng lưới của họ, quyết định nơi cần bổ sung hoặc nâng cấp cơ sở hạ tầng. Nhưng dữ liệu còn có nhiều ứng dụng tiềm năng khác nữa. Các nhà sản xuất thiết bị cầm tay có thể sử dụng nó để tìm hiểu những gì ảnh hưởng đến cường độ tín hiệu, ví dụ để nâng cao chất lượng tiếp nhận tín hiệu cho các thiết bị của họ. Các nhà khai thác điện thoại di động từ lâu đã không muốn kiếm tiền từ thông tin này vì sợ vi phạm các quy định bảo vệ quyền riêng tư. Nhưng họ bắt đầu mềm dẻo hơn

trong lập trường khi dữ liệu được xem như một nguồn thu nhập tiềm năng. Năm 2012, công ty Telefonica thậm chí còn lập ra một công ty riêng biệt, gọi là Telefonica Digital Insights, để bán dữ liệu vị trí thuê bao ẩn danh cho các nhà bán lẻ và những đối tượng khác.

DỮ LIỆU TÁI TỔ HỢP

Đôi khi giá trị tiềm ẩn chỉ có thể được giải phóng bằng cách kết hợp một bộ dữ liệu với một bộ khác, thậm chí hoàn toàn khác.

Chúng ta có thể sáng tạo bằng cách trộn lẫn dữ liệu theo những cách mới. Một ví dụ để thấy cách này vận hành như thế nào là một nghiên cứu thông minh được công bố năm 2011 để xem liệu điện thoại di động có làm tăng nguy cơ ung thư. Với khoảng sáu tỷ điện thoại di động trên thế giới, gần như một máy cho mỗi người trên trái đất, câu hỏi này là rất quan trọng. Nhiều nghiên cứu đã cố tìm kiếm một liên kết, nhưng đều gặp trở ngại do có nhiều thiếu sót. Các cỡ mẫu là quá nhỏ, hoặc những khoảng thời gian họ đề cập là quá ngắn, hoặc họ đã dựa trên dữ liệu tự báo cáo mang đầy lỗi. Tuy nhiên, một nhóm các nhà nghiên cứu tại Hiệp hội Ung thư Đan Mạch đã phát minh ra một cách tiếp cận thú vị dựa trên dữ liệu đã thu thập được trước đó.

Dữ liệu về tất cả các thuê bao từ khi có điện thoại di động ở Đan Mạch được thu thập từ các nhà khai thác di động. Nghiên cứu đã khảo sát những người có điện thoại di động từ năm 1987 đến 1995, loại trừ các thuê bao của công ty và những người không có sẵn dữ liệu kinh tế xã hội. Tổng cộng có 358.403 người. Quốc gia này cũng duy trì một cơ sở dữ liệu toàn quốc của tất cả các bệnh nhân ung thư, trong đó có 10.729 người có khối u ở hệ thống thần kinh trung ương trong những năm từ 1990 đến 2007. Nghiên cứu cũng sử dụng một cơ sở dữ liệu toàn quốc với thông tin về cấp giáo dục cao nhất và thu nhập của mỗi người dân Đan Mạch. Sau khi kết hợp ba bộ dữ liệu, các nhà nghiên cứu xem xét

liệu người sử dụng điện thoại di động có tỷ lệ ung thư cao hơn so với những người không sử dụng hay không. Và giữa các thuê bao, liệu những người đã sở hữu một điện thoại di động trong một thời gian dài hơn có nhiều khả năng bị ung thư hơn không?

Dù nghiên cứu này ở quy mô lớn, dữ liệu thu được không hề lộn xộn hoặc thiếu chính xác: các bộ dữ liệu đòi hỏi những tiêu chuẩn chất lượng khắt khe cho các mục đích y tế, thương mại hoặc nhân khẩu học. Thông tin được thu thập không theo những cách có thể tạo ra định kiến liên quan đến chủ đề của nghiên cứu. Thật ra, dữ liệu đã có từ nhiều năm trước, vì những lý do không hề liên quan tới nghiên cứu này. Điều quan trọng nhất là nghiên cứu không dựa trên một mẫu mà trên cơ sở gần với N = tất cả: hầu hết các ca bệnh ung thư, và gần như tất cả người dùng điện thoại di động, với số lượng 3,8 triệu người và số năm sở hữu điện thoại di động. Việc nó bao gồm gần như tất cả các trường hợp nghĩa là các nhà nghiên cứu có thể kiểm soát các tiểu quần thể, chẳng hạn như những người có mức thu nhập cao.

Cuối cùng, nhóm đã không phát hiện được bất kỳ sự gia tăng nguy cơ ung thu nào liên quan với việc sử dụng điện thoại di động. Vì lý do đó, các kết quả của nghiên cứu hầu nhu không gây được tiếng vang trên các phương tiện truyền thông khi chúng được công bố vào tháng 10 năm 2011 trên tạp chí y khoa của Anh *BMJ.* Nhưng nếu một mối liên hệ được phát hiện thì nghiên cứu này hẳn sẽ xuất hiện trên trang nhất của các tờ báo khắp thế giới, và phương pháp “dữ liệu tái tổ hợp” đã nổi tiếng.

Với dữ liệu lớn, tổng thể sẽ có giá trị cao hơn các bộ phận của nó, và khi chúng ta kết hợp các tổng thể của nhiều bộ dữ liệu lại với nhau, tổng thể đó cũng là trị giá hơn các thành phần riêng lẻ. Ngày nay người dùng Internet quen thuộc với những “ứng dụng hỗn hợp” cơ bản, kết hợp hai hoặc nhiều nguồn dữ liệu theo một cách mới lạ. Ví dụ trang web bất động sản Zillow đã chèn thông tin bất động sản và giá cả lên bản đồ của các khu phố tại Hoa Kỳ.

Họ cũng xử lý hàng núi dữ liệu, chẳng hạn các giao dịch gần đây trong khu vực và chi tiết kỹ thuật của các bất động sản, để dự đoán giá trị của những ngôi nhà cụ thể trong một khu vực. Cách trình bày hình ảnh làm cho dữ liệu trở nên dễ tiếp cận hơn. Nhưng với dữ liệu lớn chúng ta còn có thể đi xa hơn nữa. Nghiên cứu về ung thư ở Đan Mạch đã cho chúng ta một gợi ý về những điều khả thi.

DỮ LIỆU MỞ RỘNG

Một cách khiến việc tái sử dụng dữ liệu dễ dàng hơn là thiết kế khả năng mở rộng cho nó ngay từ đầu, để nó phù hợp với nhiều mục đích sử dụng. Mặc dù điều này không phải luôn khả thi - bởi có thể rất lâu sau khi dữ liệu đã được thu thập người ta mới nhận ra những ứng dụng khác - vẫn có nhiều cách khuyên khích các ứng dụng khác nhau cho cùng một bộ dữ liệu. Ví dụ một số cửa hàng bán lẻ đặt các camera giám sát cửa hàng, không chỉ để phát hiện người lấy cắp đồ, mà quan trọng là để theo dõi dòng khách mua trong cửa hàng và nơi họ dừng lại nhìn ngắm. Các nhà bán lẻ có thể sử dụng nhóm thông tin này để thiết kế cửa hàng cũng như để đánh giá hiệu quả của các chiến dịch tiếp thị. Trước đó, camera chỉ phục vụ mục tiêu an ninh. Bây giờ chúng được xem là khoản đầu tư có thể làm tăng doanh thu.

Một trong những công ty giỏi nhất trong việc thu thập dữ liệu, đồng thời tính đến khả năng mở rộng, đương nhiên chính là Google. Những chiếc xe Street View vốn gây tranh cãi đã đi khắp nơi chụp ảnh nhà ở và đường giao thông, nhưng cũng ngấu nghiến dữ liệu GPS, kiểm tra thông tin bản đồ, thậm chí lấy các tên mạng wifi (và cả nội dung truyền tải trên các mạng wifi mở, có lẽ một cách bất hợp pháp). Chỉ một chuyến đi của Google Street View đã tích lũy được vô số dòng dữ liệu rời rạc ở mọi thời điểm. Khả năng mở rộng xuất hiện bởi vì Google dùng các dữ liệu không chỉ cho ứng dụng chính mà còn cho rất nhiều các ứng dụng phụ. Ví dụ dữ liệu GPS thu thập được đã cải thiện dịch vụ

bản đồ của họ và là phần không thể thiếu cho hoạt động của Google Street View.

Chi phí phát sinh để thu thập nhiều dòng hoặc nhiều điểm dữ liệu hơn trong mỗi dòng thường thấp. Vì vậy, rõ ràng là thu thập càng nhiều dữ liệu càng tốt, cũng như cần làm cho dữ liệu có thể được mở rộng bằng cách xem xét tiềm năng của các ứng dụng phụ ngay từ đầu. Điều này làm tăng giá trị lựa chọn của dữ liệu. Vấn đề là tìm các "ích lợi kép" - nghĩa là một bộ dữ liệu đơn nhất có thể được sử dụng trong nhiều trường hợp nếu nó được thu thập theo một cách nhất định. Nhờ đó, dữ liệu có thể thực thi nhiều nhiệm vụ cùng lúc.

GIẢM GIÁ TRỊ CỦA DỮ LIỆU

Khi chi phí lưu trữ dữ liệu kỹ thuật số đã giảm mạnh, các doanh nghiệp có động lực kinh tế mạnh mẽ trong việc giữ lại dữ liệu để tái sử dụng cho cùng mục đích hoặc cho những mục đích tương tự khác. Nhưng có một giới hạn cho tính hữu dụng của nó.

Ví dụ các công ty như NetAix và Amazon dựa vào các giao dịch của khách hàng và các đánh giá để đưa ra khuyến nghị cho các sản phẩm mới, do vậy họ có thể chấp nhận sử dụng các hồ sơ nhiều lần cho nhiều năm. Với ý nghĩ đó, người ta có thể tranh luận rằng khi không bị hạn chế bởi các giới hạn pháp lý như luật bảo vệ quyền riêng tư, công ty nên sử dụng các hồ sơ kỹ thuật số mãi mãi, hoặc ít nhất là khi vẫn còn hiệu quả về mặt kinh tế. Tuy nhiên, thực tế lại không đơn giản như vậy.

Hầu hết dữ liệu đều bị mất một phần tính hữu ích của nó theo thời gian. Trong những hoàn cảnh như vậy, việc tiếp tục dựa vào dữ liệu cũ không chỉ thất bại trong việc gia tăng giá trị, nó còn thực sự phá hủy giá trị của dữ liệu mới hơn. Hãy chọn một cuốn sách bạn mua mười năm trước từ Amazon mà nó có thể không còn phản ánh các sở thích của bạn nữa. Nếu Amazon sử dụng hồ

sơ mua hàng cũ cả chục năm để giới thiệu các cuốn sách khác thì ít có khả năng bạn sẽ mua chúng - hoặc thậm chí thèm để tâm tới các khuyến nghị tiếp theo mà trang web cung cấp. Khi các khuyến nghị của Amazon dựa trên cả thông tin lỗi thời lẫn thông tin gần đây hơn vẫn còn giá trị, sự hiện diện của các dữ liệu cũ sẽ làm giảm giá trị của các dữ liệu mới hơn.

Vì vậy, công ty vẫn sử dụng dữ liệu chỉ khi nó vẫn còn có hiệu quả. Amazon cần liên tục chăm chút kho tàng dữ liệu và xóa bỏ các thông tin đã mất giá trị. Khó khăn nằm ở chỗ biết được dữ liệu nào không còn hữu ích nữa. Nếu chỉ ra quyết định căn cứ vào thời gian thì hiếm khi thỏa đáng. Do đó, Amazon và những công ty khác đã xây dựng những mô hình phức tạp để giúp họ tách biệt dữ liệu hữu ích với dữ liệu không liên quan. Ví dụ nếu một khách hàng xem hoặc mua một cuốn sách được đề nghị dựa trên một lần mua trước, thì công ty thương mại điện tử có thể suy ra rằng giao dịch cũ vẫn còn thể hiện cho những sở thích hiện tại của khách hàng. Bằng cách đó họ có thể chấm điểm cho tính hữu dụng của dữ liệu cũ, và nhờ đó lập ra mô hình "tỷ lệ khấu hao" chính xác hơn cho các thông tin.

Không phải tất cả dữ liệu đều mất giá trị với cùng một tốc độ hoặc theo cùng một cách. Điều này giải thích lý do một số công ty tin rằng họ cần lưu trữ dữ liệu càng lâu càng tốt, ngay cả khi các cơ quan quản lý hoặc công chúng muốn nó được xóa đi hoặc làm ẩn danh sau một thời gian. Ví dụ Google từ lâu đã phản đối các yêu cầu xóa địa chỉ giao thức Internet đầy đủ của người sử dụng từ các truy vấn tìm kiếm cũ. (Thay vào đó nó chỉ xóa chữ số cuối cùng sau chín tháng để làm ẩn danh một phần các truy vấn. Như vậy, công ty vẫn có thể so sánh dữ liệu năm này qua năm khác, chẳng hạn các lệnh tìm kiếm về mua sắm dịp lễ - nhưng chỉ trên cơ sở khu vực, chứ không xuống tới từng cá nhân.) Ngoài ra, việc biết vị trí của người tìm kiếm có thể giúp cải thiện tính xác đáng của các kết quả. Ví dụ nếu nhiều người ở New York tìm kiếm và mở trang web về Thổ Nhĩ Kỳ, thuật toán sẽ xếp hạng các

trang này cao hơn cho những người khác ở New York. Ngay cả khi giá trị của dữ liệu giảm đối với một số mục đích của nó, giá trị tương lai của nó có thể vẫn còn lớn.

Giá trị của dữ liệu xả

Tái sử dụng dữ liệu đôi khi có thể ở một hình thức thông minh và ẩn. Các công ty web có thể thu thập dữ liệu trên tất cả những điều mà người sử dụng thực hiện, và sau đó xử lý mỗi tương tác riêng biệt như một chỉ báo có vai trò là thông tin phản hồi để phục vụ việc cá nhân hóa trang web, cải thiện dịch vụ, hoặc tạo ra một sản phẩm kỹ thuật số hoàn toàn mới. Chúng ta sẽ thấy một minh họa sinh động về điều này trong câu chuyện về hai bộ kiểm tra chính tả.

Trong suốt hai mươi năm qua, Microsoft đã phát triển một bộ kiểm tra chính tả mạnh cho phần mềm Word. Nó so sánh một từ điển thường xuyên được cập nhật của các từ viết đúng chính tả với dòng các ký tự người sử dụng gõ vào. Từ điển lập danh sách những từ đã được biết đến, và hệ thống sẽ xem các biến thể gần đúng nhưng không có trong từ điển là lỗi chính tả để sau đó sửa. Do sẽ phải tiêu tốn nhiều công sức để sưu tập và cập nhật từ điển, bộ kiểm tra chính tả của Microsoft Word chỉ có cho những ngôn ngữ phổ biến nhất. Nó tiêu tốn của công ty hàng triệu đôla để tạo ra và duy trì sản phẩm.

Bây giờ hãy sang Google. Họ được cho là có bộ kiểm tra chính tả hoàn thiện nhất thế giới, về cơ bản là cho tất cả các ngôn ngữ được sử dụng. Hệ thống liên tục cải thiện và bổ sung thêm những từ mới - kết quả ngẫu nhiên của việc mọi người sử dụng công cụ tìm kiếm mỗi ngày. Gõ nhầm “iPad”? Đã có trong dữ liệu. “Obamacare”? Nó biết luôn rồi.

Hơn nữa, Google dường như có được bộ kiểm tra chính tả mà chẳng tốn phí, do tái sử dụng các lỗi chính tả được gõ vào công

cụ tìm kiếm của ba tỷ yêu cầu mà nó xử lý mỗi ngày. Một vòng phản hồi thông minh dạy cho hệ thống từ nào là từ người sử dụng thực sự muốn gõ vào. Người sử dụng đôi khi "nói" một cách rõ ràng cho Google câu trả lời khi nó đặt ra câu hỏi ở trên cùng của trang kết quả - ví dụ "Ý của bạn là *epidemioiogy*?" - bằng cách nhấp vào đó để bắt đầu một lệnh tìm kiếm mới với từ khóa đúng. Hoặc trang web mà người dùng muốn nhắm tới sẽ giả định việc viết đúng chính tả, có thể vì như vậy sẽ tương hợp hơn so với từ khóa viết sai. (Điều này là quan trọng hơn nhiều người tưởng: Khi bộ kiểm tra chính tả của Google được liên tục cải tiến, người ta không cần gõ các từ khóa tìm kiếm của họ một cách chính xác nữa, bởi Google vẫn có thể xử lý chúng được.)

Hệ thống kiểm tra chính tả của Google cho thấy dữ liệu "xấu", "không đúng", hoặc "khiếm khuyết" vẫn có thể rất hữu ích. Điều thú vị là Google không phải là nơi đầu tiên có ý tưởng này. Khoảng năm 2000 Yahoo đã nhìn thấy khả năng tạo ra một bộ kiểm tra chính tả từ các truy vấn gõ sai của người sử dụng. Nhưng ý tưởng này chẳng đi được tới đâu. Dữ liệu câu hỏi tìm kiếm cũ đã bị xử lý chủ yếu như là rác. Tương tự như vậy, Infoseek và Alta Vista, những công cụ tìm kiếm phổ biến sớm hơn, đều có cơ sở dữ liệu toàn diện nhất thế giới về các từ viết sai chính tả khi đó, nhưng họ đã không đánh giá cao giá trị của chúng. Các hệ thống của họ, trong một quá trình ẩn đối với người sử dụng, đã xem những từ viết sai như "những từ có liên quan" và vẫn tiến hành cuộc tìm kiếm. Nhưng cuộc tìm kiếm đó được dựa trên các từ điển nói rõ ràng với hệ thống những gì là đúng, chứ không dựa trên những điều sống động, hiện hữu của việc tương tác với người dùng.

Chỉ mỗi Google nhận ra những mảnh vụn của mối tương tác với người dùng là bụi vàng thực sự, có thể được thu thập lại và đúc thành một phôi sáng bóng. Một trong những kỹ sư hàng đầu của Google ước tính rằng bộ kiểm tra chính tả của nó thực hiện tốt hơn so với của Microsoft ở mức độ rất cao (mặc dù khi được chất

vấn, ông thừa nhận đã không đo lường điều này một cách đáng tin cậy). Và ông chế giễu ý kiến cho rằng nó được phát triển "miễn phí". Có thể nguyên liệu thô - lỗi chính tả - tự đến mà không cần một chi phí trực tiếp nào, nhưng Google nhiều khả năng đã chi nhiều hơn hẳn so với Microsoft để phát triển hệ thống, ông thừa nhận với một nụ cười sảng khoái.

Các phương pháp tiếp cận khác nhau của hai công ty là vô cùng đáng chú ý. Microsoft chỉ nhìn thấy giá trị của việc kiểm tra chính tả cho một mục đích: xử lý từ. Google lại hiểu được ích lợi sâu hơn của nó. Google không chỉ sử dụng các lỗi chính tả nhằm phát triển bộ kiểm tra chính tả tốt nhất và được cập nhật tốt nhất thế giới để cải thiện việc tìm kiếm, mà nó còn áp dụng hệ thống vào nhiều dịch vụ khác, chẳng hạn như tính năng "tự động hoàn chỉnh" trong tìm kiếm, Gmail, Google Docs, và thậm chí cả hệ thống dịch thuật của mình.

Một thuật ngữ nghệ thuật đã xuất hiện để mô tả dấu vết kỹ thuật số mà người sử dụng để lại: "dữ liệu xả". Nó đề cập đến dữ liệu được tạo ra như một sản phẩm phụ của các hành vi và các chuyển động của con người trong thế giới. Với Internet, nó mô tả những tương tác trực tuyến của người sử dụng: nơi họ nhấp chuột, họ xem một trang bao lâu, nơi con trỏ chuột qua lại, những gì họ nhập từ bàn phím, và nhiều nữa. Nhiều công ty thiết kế hệ thống của họ để có thể thu hoạch được dữ liệu xả và tái chế, để cải thiện một dịch vụ hiện có hoặc phát triển những dịch vụ mới. Google là người dẫn đầu không thể tranh cãi. Nó áp dụng nguyên tắc đệ quy "học hỏi từ dữ liệu" cho nhiều dịch vụ của mình. Mọi hành động người dùng thực hiện được xem là một tín hiệu để phân tích và đưa trở lại vào hệ thống.

Ví dụ Google nhận thức được một cách sâu sắc việc bao nhiêu lần người dùng tìm kiếm một từ khóa cũng như những từ liên quan, và mức độ thường xuyên họ bấm vào một liên kết nhưng sau đó quay trở lại trang tìm kiếm vì không hài lòng với những gì họ

tìm thấy, để tìm kiếm một lần nữa. Nó biết liệu họ đang bấm vào liên kết thứ tám trên trang đầu tiên hay liên kết đầu tiên trên trang thứ tám - hay họ đã từ bỏ hoàn toàn việc tìm kiếm. Google có thể không phải là công ty đầu tiên có cái nhìn sâu sắc này, nhưng là công ty thực hiện điều này với hiệu quả đặc biệt xuất sắc.

Thông tin này rất có giá trị. Nếu nhiều người dùng có xu hướng bấm vào kết quả tìm kiếm ở dưới cùng của hang kết quả, điều này cho thấy nó phù hợp hơn những kết quả được xếp trên, và thuật toán xếp hạng của Google sẽ biết để tự động đặt nó lên cao hơn trong những lần tìm kiếm tiếp theo. (Và nó thực hiện điều này cho cả những quảng cáo.) "Chúng tôi thích học hỏi từ những tập hợp dữ liệu lớn, 'ồn ào'", một chuyên gia của Google nhận xét.

Dữ liệu xả là cơ chế đằng sau rất nhiều dịch vụ như nhận dạng giọng nói, lọc thư rác, dịch ngôn ngữ, và nhiều nữa. Khi người sử dụng chỉ cho một chương trình nhận dạng giọng nói rằng nó đã hiểu lầm những gì họ nói, họ thực chất đã "huấn luyện" hệ thống để nó tốt hơn. Nhiều doanh nghiệp đang bắt đầu thiết kế hệ thống của họ để thu thập và sử dụng thông tin theo cách này. Trong những ngày đầu của Facebook, "nhà khoa học dữ liệu" đầu tiên của công ty, Jeff Hammerbacher (và là một trong số những người đặt ra thuật ngữ này), đã khảo sát kho tàng phong phú của dữ liệu xả. Ông và nhóm nghiên cứu phát hiện ra rằng một yếu tố dự báo lớn về việc người dùng sẽ thực hiện một hành động (đăng nội dung, nhấp vào một biểu tượng...) là liệu họ có nhìn thấy bạn bè của mình làm điều tương tự hay không. Vì vậy, Facebook đã thiết kế lại hệ thống để chú trọng nhiều hơn vào việc khiến cho các hoạt động của bạn bè có thể được nhìn thấy rõ hơn, tạo ra một vòng xoắn phát triển của những đóng góp mới cho trang web.

Ý tưởng này đang lan rộng vượt ra ngoài lĩnh vực Internet tới bất kỳ công ty nào thu thập thông tin phản hồi của người dùng. Ví

dụ những thiết bị đọc sách điện tử (e-book) nắm bắt số lượng lớn dữ liệu về sở thích và thói quen văn học của người sử dụng chúng: họ cần bao lâu để đọc một trang hoặc đoạn, nơi họ đọc, họ lật trang chỉ để lướt qua hoặc gấp cuốn sách lại mãi mãi. Các thiết bị ghi lại mỗi khi người sử dụng đánh dấu một đoạn hoặc ghi chú ở bên lề. Khả năng thu thập loại thông tin này sẽ biến việc đọc, lâu nay là một hành động đơn độc, thành một loại trải nghiệm chung.

Một khi đã được tổng hợp, dữ liệu xả có thể cho các nhà xuất bản và tác giả biết những điều mà họ chưa hề được biết trước đây một cách định lượng: các cảm giác thích, không thích, và mô thức đọc của mọi người. Thông tin này rất có giá trị về thương mại. Có thể hình dung các công ty sách điện tử bán nó cho các nhà xuất bản để cải tiến nội dung và cấu trúc của các cuốn sách. Ví dụ việc phân tích dữ liệu từ thiết bị đọc sách điện tử Nook của Barnes & Noble cho thấy khi đọc một tác phẩm dày thuộc thể loại sách kiến thức, người ta thường bỏ ngang khi chỉ mới đọc được một nửa. Phát hiện này đã tạo cảm hứng cho công ty cho ra đời loạt sách được gọi là “Nook Snaps”: những tác phẩm ngắn về các chủ đề thời sự như y tế và các vấn đề đương đại.

Hoặc hãy xem xét các chương trình đào tạo trực tuyến như Udacity, Coursera, và edX. Chúng theo dõi các tương tác web của học sinh để xem điều gì là tốt nhất về mặt sư phạm. Các lớp học có quy mô hàng chục ngàn học sinh, tạo ra lượng dữ liệu vô cùng lớn. Các giáo sư nay có thể biết khi một tỷ lệ lớn sinh viên xem lại một phân đoạn của một bài giảng, và điều đó có thể do họ chưa rõ về một điểm nào đó. Khi giảng dạy một lớp của Coursera về máy tính, giáo sư Andrew Ng của Stanford nhận thấy khoảng 2.000 sinh viên làm sai một câu hỏi trong bài tập về nhà - nhưng đưa ra chính xác cùng một câu trả lời sai. Rõ ràng, tất cả họ đã mắc cùng một lỗi. Nhưng lỗi đó là gì?

Sau khi điều tra, ông phát hiện ra rằng họ đã đảo ngược hai

phương trình đại số trong một thuật toán. Vậy nên từ bây giờ, khi những sinh viên khác mắc cùng một lỗi, hệ thống không chỉ đơn giản nói họ sai, mà còn cho họ một gợi ý để kiểm tra lại phép tính. Hệ thống này cũng áp dụng dữ liệu lớn, bằng cách phân tích mỗi bài viết trong diễn đàn mà sinh viên đọc và họ hoàn thành bài tập về nhà một cách chính xác hay không để dự đoán xác suất mà một sinh viên đã đọc một bài viết nhất định sẽ đưa ra kết quả đúng, như một cách để xác định những bài viết nào trên diễn đàn là hữu ích nhất cho sinh viên đọc. Đây là những điều hoàn toàn không thể biết được trước đây, và có thể làm thay đổi việc dạy và học mãi mãi. Dữ liệu xả có thể là một lợi thế cạnh tranh rất lớn cho các công ty.

Nó cũng có thể trở thành một rào cản mạnh mẽ để ngăn đối thủ mới xuất hiện. Nếu một công ty vừa thành lập tạo ra một trang web thương mại điện tử, mạng xã hội, hay công cụ tìm kiếm tốt hơn rất nhiều so với các công ty hàng đầu hiện nay như Amazon, Google, hay Facebook, nó sẽ gặp khó khăn khi cạnh tranh, không chỉ đơn giản vì những hiệu ứng của kinh tế quy mô lớn và mạng lưới hoặc thương hiệu, mà còn vì phần lớn hiệu suất của những công ty hàng đầu này là từ dữ liệu xả họ thu thập từ các tương tác của khách hàng và kết hợp trở lại vào dịch vụ. Liệu một dịch vụ web đào tạo trực tuyến mới có đủ sức cạnh tranh với một địch thủ đã có một lượng khổng lồ dữ liệu để giúp nó tìm hiểu được những gì sẽ hoạt động hiệu quả nhất?

Giá trị của dữ liệu mở

Ngày nay chúng ta dễ nghĩ các trang web như Google và Amazon là những nhà tiên phong của dữ liệu lớn, nhưng tất nhiên các chính phủ mới là những người thu lượm thông tin gốc trên quy mô lớn, và họ sẽ không kém cạnh bất kỳ doanh nghiệp tư nhân nào về khối lượng lớn dữ liệu mà họ kiểm soát. Một sự khác biệt với các chủ sở hữu dữ liệu trong khu vực tư nhân là các chính phủ thường có thể bắt buộc mọi người cung cấp thông tin, chứ

không phải thuyết phục họ làm như vậy hoặc phải trả cho họ một cái gì đó để đổi lại. Do đó, chính phủ sẽ vẫn tiếp tục tích lũy được những kho tàng lớn dữ liệu.

Những bài học của dữ liệu lớn áp dụng cho khu vực công cũng giống như cho các cơ sở thương mại: giá trị dữ liệu của chính phủ là tiềm ẩn và đòi hỏi việc phân tích sáng tạo để được khai thông. Nhưng bất chấp vị trí đặc biệt của mình trong việc nắm bắt thông tin, các chính phủ thường không mấy hiệu quả trong việc khai thác nó. Gần đây, một ý tưởng được nhiều người xem như cách tốt nhất để tận dụng các giá trị của dữ liệu từ chính phủ là để cho khu vực tư nhân và công chúng nói chung truy cập dữ liệu và thử nghiệm. Còn có một nguyên lý nữa ở phía sau. Khi nhà nước tập hợp dữ liệu, họ làm việc đó thay mặt cho các công dân, và do đó nhà nước phải cung cấp quyền truy cập cho đại chúng (ngoại trừ một số ít trường hợp, chẳng hạn như khi làm như vậy có thể gây tổn hại cho an ninh quốc gia hoặc các quyền riêng tư của những người khác).

Ý tưởng này đã dẫn đến vô số những sáng kiến "dữ liệu chính phủ mở" trên toàn cầu. Cho rằng chính phủ chỉ canh giữ các thông tin mà họ thu thập, còn khu vực tư nhân và đại chúng sẽ sáng tạo hơn, những người ủng hộ dữ liệu mở kêu gọi các cơ quan chính phủ công khai dữ liệu cho các mục đích dân sự và thương mại. Để làm được việc này, tất nhiên, dữ liệu phải ở một dạng chuẩn hóa, máy có thể đọc được để dễ dàng xử lý. Nếu không, các thông tin chỉ là công khai trên danh nghĩa.

Ý tưởng về dữ liệu mở của chính phủ được thúc đẩy mạnh khi Tổng thống Barack Obama, vào ngày làm việc đầu tiên của ông tại Nhà Trắng, 21 tháng 1 năm 2009, ban hành một biên bản ghi nhớ của Tổng thống ra lệnh cho người đứng đầu các cơ quan của liên bang phải công bố càng nhiều dữ liệu càng tốt. "Đối mặt với sự nghi ngờ, việc mở cửa sẽ thắng thế", ông chỉ thị. Đó là một tuyên bố đáng chú ý, đặc biệt khi so sánh với người tiền nhiệm

đã chỉ thị các cơ quan làm chính xác điều ngược lại. Lệnh của Obama thúc đẩy việc tạo ra trang web data.gov, một kho lưu trữ mở cho phép truy cập đến thông tin từ chính phủ liên bang. Trang web nhanh chóng phát triển từ 47 bộ dữ liệu trong năm 2009 lên gần 450.000 bộ dữ liệu bao gồm 172 cơ quan vào dịp kỷ niệm ba năm hoạt động, tháng 7 năm 2012.

Thậm chí ở nước Anh bảo thủ, nơi rất nhiều thông tin chính phủ đã bị khóa bởi Luật bản quyền Crown, đồng thời để được cấp giấy phép sử dụng sẽ rất khó khăn và tốn kém (chẳng hạn như mã bưu chính cho các công ty thương mại điện tử), cũng đã có tiến bộ đáng kể. Chính phủ Anh đã ban hành quy định khuyến khích thông tin mở và hỗ trợ việc thành lập một Viện Dữ liệu Mở, do Tim Berners-Lee, người phát minh ra World Wide Web, đồng lãnh đạo để thúc đẩy những ứng dụng mới mẻ của dữ liệu mở và những cách thức để giải phóng nó khỏi sự kiềm kẹp của nhà nước.

Liên minh châu Âu cũng công bố những sáng kiến dữ liệu mở có thể sớm trở thành những sáng kiến của châu lục. Các nước khác, như Úc, Brazil, Chile, và Kenya, đã ban hành và thực hiện các chiến lược dữ liệu mở. Bên dưới cấp quốc gia, một số lượng ngày càng tăng các thành phố và đô thị trên thế giới cũng chấp nhận mở dữ liệu. Các tổ chức quốc tế như Ngân hàng Thế giới đã mở cửa hàng trăm bộ dữ liệu về các chỉ tiêu kinh tế và xã hội mà trước đây đã bị giới hạn.

Song song đó, cộng đồng các nhà phát triển web và các nhà tư tưởng nhìn xa trông rộng đã được hình thành để tìm ra cách thu được nhiều nhất từ dữ liệu, ví dụ như Code for America và Quỹ Sunlight tại Hoa Kỳ, hoặc Quỹ Tri thức Mở tại Anh. Một ví dụ sớm về các khả năng của dữ liệu mở xuất phát từ trang web FlyOnTime.us. Khách truy cập vào trang web có thể tương tác để tìm hiểu (trong số nhiều mối tương quan khác) khả năng thời tiết xấu sẽ trì hoãn các chuyến bay tại một sân bay cụ thể. Trang

web kết hợp chuyến bay và thông tin thời tiết từ những nguồn số liệu chính thức được truy cập tự do qua Internet. Nó được phát triển bởi những người ủng hộ dữ-liệu-mở để biểu lộ sự hữu ích của thông tin tích lũy được của chính phủ liên bang. Ngay cả phần mềm của trang web cũng là mã nguồn mở, để những người khác có thể học hỏi từ nó và tái sử dụng nó.

FlyOnTime.us để cho dữ liệu tự *nói,* và nó thường *nói* những điều đáng ngạc nhiên. Người ta có thể thấy với những chuyên bay từ Boston đi sân bay LaGuardia New York, du khách cần chuẩn bị cho sự chậm trễ vì sương mù với thời gian dài gấp đôi so với vì tuyết. Điều này có lẽ không phải là thứ hầu hết mọi người có thể đoán được khi ngồi ở phòng chờ khởi hành; tuyết có vẻ là một nguyên nhân nghiêm trọng hơn gây chậm trễ. Nhưng đây là loại hiểu biết mà dữ liệu lớn có thể mang lại, qua khảo sát dữ liệu lịch sử các vụ trễ chuyến bay của Cục Giao thông Vận tải, thông tin sân bay hiện tại từ Cục Hàng không Liên bang, cùng với dự báo thời tiết từ Cơ quan Đại dương và Khí quyển Quốc gia và các điều kiện thời gian thực từ Cục Thời tiết Quốc gia. FlyOnTime.us cho thấy rằng một thực thể không hề thu thập hay kiểm soát dòng chảy thông tin, giống như một công cụ tìm kiếm hay nhà bán lẻ lớn, có thể vẫn nhận được và sử dụng dữ liệu để tạo ra giá trị như thế nào.

Định giá sự vô giá

Dù mở cho công chúng hay khóa kín trong hầm của công ty, giá trị của dữ liệu rất khó để đo lường. Hãy xem xét các sự kiện của ngày thứ Sáu, 18 tháng 5 năm 2012. Vào ngày đó, người sáng lập Facebook Mark Zuckerberg, 28 tuổi, đã rung chuông một cách tượng trưng từ trụ sở chính của công ty tại Menlo Park, California để mở đầu phiên giao dịch của chứng khoán NASDAQ. Mạng xã hội lớn nhất thế giới - tự hào vì có khoảng một phần mười dân số hành tinh là thành viên lúc đó - bắt đầu cuộc đời mới của mình như một công ty đại chúng, cổ phiếu ngay lập tức

tăng 11 phần trăm, giống như nhiều cổ phiếu công nghệ mới trong ngày giao dịch đầu tiên. Tuy nhiên, sau đó một điều kỳ lạ đã xảy ra. Cổ phiếu của Facebook bắt đầu rơi. Xu hướng không thay đổi khi một trục trặc kỹ thuật với máy tính của NASDAQ đã tạm thời dừng giao dịch. Một vấn đề lớn hơn đang xảy ra. Cảm thấy lo ngại, các nhà bảo lãnh phát hành chứng khoán, dẫn đầu là Morgan Stanley, đã thực sự nhảy vào hỗ trợ để giữ cổ phiếu ở trên giá phát hành.

Buổi tối hôm trước, các ngân hàng của Facebook đã định giá công ty ở mức $38 một cổ phiếu, và công ty được định giá tương đương 104 tỷ đôla. (Như vậy là xấp xỉ mức vốn hóa thị trường của Boeing, General Motors, và Dell Computers cộng lại.) Facebook thực sự có giá trị bao nhiêu? Trong báo cáo tài chính đã được kiểm toán cho năm 2011, cơ sở để các nhà đầu tư định giá công ty, Facebook công bố tài sản là $6,3 tỷ. Đó là đại diện cho giá trị của phần cứng máy tính, thiết bị văn phòng, và các công cụ vật lý khác. Đối với giá trị sổ sách trên các kho tàng lớn thông tin mà Facebook cất giữ thì sao? Về cơ bản là bằng không. Nó không được tính vào, mặc dù công ty này gần như không có gì *ngoài* dữ liệu.

Tình hình còn trở nên kỳ quặc hơn. Doug Laney, phó chủ tịch nghiên cứu của công ty nghiên cứu thị trường Gartner, phân tích các số liệu trong giai đoạn trước khi phát hành lần đầu ra công chúng (IPO) và cho rằng Facebook đã thu thập được 2,1 nghìn tỷ mục "nội dung có thể định giá" từ năm 2009 đến 2011, ví dụ như các nội dung "thích", các tư liệu đăng tải, và các ý kiến. So sánh với việc định giá IPO thì điều này có nghĩa là mỗi mục, được xem như một điểm dữ liệu rời rạc, có giá trị khoảng 5 cent. Theo một cách nhìn khác, mỗi người sử dụng Facebook có giá trị khoảng $100, bởi vì người sử dụng là nguồn gốc của các thông tin mà Facebook thu thập.

Làm thế nào để giải thích sự chênh lệch lớn giữa giá trị của

Facebook theo các chuẩn mực kế toán ($6,3 tỷ) và những gì thị trường ban đầu định giá nó ($104 tỷ)? Không có cách nào đủ tốt để làm việc này. Thay vào đó, người ta thống nhất phương pháp hiện hành xác định giá trị của công ty bằng cách nhìn vào "giá trị sổ sách" của nó (nghĩa là chủ yếu gồm giá trị tiền mặt và các tài sản vật chất). Cách này không còn phản ánh đầy đủ giá trị thực sự. Thật ra, khoảng cách giữa giá trị sổ sách và "giá trị thị trường" - những gì công ty sẽ thu được trên thị trường chứng khoán hoặc nếu nó được mua toàn bộ - đã tăng qua nhiều thập kỷ. Thượng viện Mỹ thậm chí đã có những buổi điều trần trong năm 2000 về hiện đại hóa các quy định báo cáo tài chính, những thứ được xây dựng từ những năm 1930 khi các doanh nghiệp dựa trên thông tin hầu như không tồn tại. Vấn đề này ảnh hưởng đến nhiều thứ chứ không chỉ bảng cân đối tài chính của công ty: việc không thể đánh giá đúng giá trị của công ty làm phát sinh rủi ro trong kinh doanh và gây bất ổn trên thị trường.

Sự khác biệt giữa giá trị sổ sách của công ty và giá trị thị trường của nó được ghi nhận là "tài sản vô hình". Nó đã tăng từ khoảng 40 phần trăm giá trị của các công ty giao dịch công khai ở Hoa Kỳ vào giữa những năm 1980 lên đến ba phần tư giá trị của chúng vào đầu thiên niên kỷ mới. Đây là sự phân kỳ lớn. Những tài sản vô hình này được xem là bao gồm thương hiệu, tài năng, và chiến lược - bất cứ thứ gì phi vật chất và là thành phần của hệ thống tài chính kế toán hình thức. Và càng ngày, tài sản vô hình càng gắn với dữ liệu mà công ty nắm giữ và sử dụng.

Cuối cùng, điều này cho thấy hiện nay không có cách rõ ràng để xác định giá trị dữ liệu. Ngày đầu giao dịch cổ phiếu Facebook, khoảng cách giữa tài sản chính thức và giá trị vô hình không được ghi lại của nó là gần $100 tỷ. Đúng là khôi hài. Tuy nhiên, khoảng cách này phải và sẽ khép lại bởi các công ty sẽ tìm cách ghi nhận giá trị tài sản dữ liệu của họ trong bảng cân đối.

Những bước đi ban đầu theo hướng này đang được tiến hành.

Một giám đốc điều hành cao cấp tại một trong những nhà khai thác mạng vô tuyến lớn nhất nước Mỹ cho biết các nhà khai thác đã nhận ra giá trị to lớn của dữ liệu và nghiên cứu liệu có nên xem nó như một tài sản của công ty trên phương diện kế toán chính thức. Nhưng ngay khi các luật sư của công ty nghe nói về sáng kiến này, họ đã dừng nó lại. Đưa dữ liệu lên sổ sách có thể khiến công ty phải chịu trách nhiệm về mặt pháp lý với nó, các cây đại thụ trong ngành luật lập luận, và họ cho rằng đó chẳng phải một ý tưởng hay ho gì.

Trong khi đó, các nhà đầu tư cũng bắt đầu chú ý đến giá trị tương lai của dữ liệu. Giá cổ phiếu có thể tăng lên với các công ty nào có dữ liệu hoặc có thể thu thập dữ liệu một cách dễ dàng, trong khi những công ty khác ở các vị trí kém may mắn hơn có thể thấy giá thị trường của họ co lại. Dữ liệu không nhất thiết phải chính thức xuất hiện trên các bảng cân đối để khiến điều này xảy ra. Thị trường và các nhà đầu tư sẽ đưa những tài sản vô hình này vào việc định giá - mặc dù sẽ khó khăn, như các biến động giá cổ phiếu của Facebook trong mấy tháng đầu tiên minh chứng. Nhưng khi những khó khăn về kế toán và những lo lắng về trách nhiệm được giảm bớt, gần như chắc chắn giá trị của dữ liệu sẽ hiển thị trên các bảng cân đối của công ty và trở thành một loại tài sản mới.

Vậy dữ liệu được định giá như thế nào? Việc tính toán giá trị của nó sẽ không chỉ đơn giản là cộng những gì đã đạt được từ ứng dụng chính của nó. Nếu hầu hết giá trị của dữ liệu là tiềm ẩn và có nguồn gốc từ những ứng dụng phụ chưa biết trong tương lai, cách ước lượng nó sẽ không thể ngay lập tức trở nên rõ ràng. Điều này tương tự như những khó khăn của việc định giá các yếu tố tài chính phát sinh trước khi phát triển phương trình Black-Scholes trong những năm 1970, hoặc khó khăn trong việc xác định giá trị bằng sáng chế, lĩnh vực mà các vụ bán đấu giá, trao đổi, bán hàng tư nhân, cấp phép, và rất nhiều vụ kiện tụng đang dần tạo ra một thị trường của tri thức. Ít ra, việc áp đặt một giá

cho giá trị tương lai của dữ liệu chắc chắn thể hiện một cơ hội quý báu cho lĩnh vực tài chính.

Một cách để bắt đầu là xem xét các chiến lược khác nhau mà những người sở hữu dữ liệu áp dụng để tận dụng giá trị. Khả năng rõ ràng nhất là để phục vụ cho việc ứng dụng riêng của chính công ty. Tuy nhiên, một công ty khó có khả năng phát hiện ra tất cả các giá trị tiềm ẩn của dữ liệu. Do đó với một tham vọng lớn hơn, công ty có thể cấp giấy phép sử dụng dữ liệu cho bên thứ ba. Trong thời đại dữ-liệu-lớn, nhiều chủ sở hữu dữ liệu có thể muốn lựa chọn một thỏa thuận trả một tỷ lệ phần trăm giá trị trích xuất từ dữ liệu thay vì một khoản phí cố định. Nó tương tự như các nhà xuất bản phải trả một tỷ lệ phần trăm của doanh thu từ sách, nhạc, hay phim với vai trò tiền bản quyền cho tác giả và người biểu diễn. Nó cũng giống như những thỏa thuận sở hữu trí tuệ trong công nghệ sinh học, bên cấp giấy phép có thể yêu cầu tiền bản quyền trên bất cứ phát minh tiếp theo nào xuất phát từ công nghệ của họ. Bằng cách này, tất cả các bên đều có động cơ để tối đa hóa giá trị thu được từ việc tái sử dụng dữ liệu.

Tuy nhiên, do người được cấp phép có thể thất bại trong việc tận dụng toàn bộ giá trị tương lai, chủ sở hữu dữ liệu có thể không muốn cấp quyền truy cập tài sản của họ theo kiểu độc quyền. Thay vì vậy, "dữ liệu chung chạ" có thể trở thành tiêu chuẩn. Bằng cách đó, họ có thể tự bảo hiểm cho mình.

Một số thị trường đã ra đời để thử nghiệm với những cách thức định giá dữ liệu. DataMarket, được thành lập ở Iceland vào năm 2008, cung cấp quyền truy cập tới các bộ dữ liệu miễn phí từ các nguồn khác, chẳng hạn như Liên Hiệp Quốc, Ngân hàng Thế giới, và Eurostat, và kiếm doanh thu bằng cách bán lại dữ liệu từ các nhà cung cấp thương mại như các công ty nghiên cứu thị trường. Những công ty mới thành lập khác cố gắng trở thành trung gian về thông tin, nền tảng cho các bên thứ ba chia sẻ dữ

liệu của họ, miễn phí hoặc có tính phí. Ý tưởng ở đây là để cho phép bất cứ ai bán dữ liệu có trong cơ sở dữ liệu của họ, giống như eBay cung cấp một nền tảng cho người dân bán những thứ trong gác xép của họ. Import.io khuyến khích các công ty cấp phép dữ liệu của họ, những loại dữ liệu dễ bị “cướp” từ Internet và sử dụng miễn phí. Và Factual, công ty do cựu thành viên của Google Gil Elbaz thành lập, đang cung cấp các bộ dữ liệu mà nó đã bỏ thời gian để tự sưu tập.

Microsoft cũng bước vào lĩnh vực này với Windows Azure Marketplace, tập trung vào dữ liệu chất lượng cao và giám sát những gì đang được chào bán, tương tự như cách Apple giám sát các dịch vụ trong cửa hàng ứng dụng của nó. Với tầm nhìn của Microsoft, một nhà quản lý tiếp thị làm việc trên một bảng tính Excel có thể muốn lập bảng phối hợp dữ liệu nội bộ công ty của mình với các dự báo tăng trưởng GDP từ một hãng tư vấn kinh tế. Vì vậy cô nhấp chuột để mua dữ liệu ở nơi này hoặc nơi kia, và ngay lập tức dữ liệu “chảy” vào các cột bảng tính của cô trên màn hình.

Cho đến nay vẫn chưa có thông tin về việc các mô hình định giá sẽ diễn ra như thế nào. Nhưng điều chắc chắn là nền kinh tế đang bắt đầu hình thành xung quanh dữ liệu - và nhiều người mới tham gia sẽ được hưởng lợi, trong khi một số người cũ có thể sẽ ngạc nhiên thấy luồng sinh khí mới. “Dữ liệu là một nền tảng”, theo lời của Tim O’Reilly, một nhà xuất bản công nghệ và một học giả của Thung Lũng Silicon, vì nó là một khối xây dựng cho các hàng hóa và mô hình kinh doanh mới.

Điểm mấu chốt trong giá trị của dữ liệu là tiềm năng dường như không giới hạn của nó cho tái sử dụng: giá trị tương lai. Việc thu thập thông tin tuy rất quan trọng nhưng không đủ, vì hầu hết giá trị của dữ liệu nằm ở công dụng của nó, chứ không chỉ ở chỗ sở hữu nó. Trong chương tiếp theo, chúng ta sẽ xem thật ra dữ liệu đang được sử dụng và các doanh nghiệp dữ-liệu-lớn đang

nổi lên như thế nào.

7. NHỮNG TÁC ĐỘNG

NĂM 2011 MỘT CÔNG TY MỚI RA ĐỜI Ở Seattle tên là Decide.com đã mở những cánh cửa trực tuyến của mình với tham vọng rất tuyệt vời. Nó muốn trở thành một công cụ dự-đoán-giá cho vô số các sản phẩm tiêu dùng. Nhưng nó dự định bắt đầu một cách tương đối khiêm tốn: bằng tất cả các thiết bị công nghệ cao có thể, từ điện thoại di động và TV màn hình phẳng tới máy ảnh kỹ thuật số. Những chiếc máy tính của nó bới dữ liệu từ các trang web thương mại điện tử và lùng sục trên mạng để lấy bất cứ thông tin nào về giá và sản phẩm có thể tìm thấy.

Giá cả trên mạng liên tục thay đổi suốt ngày, tự động cập nhật dựa trên vô số yếu tố phức tạp. Vì vậy, công ty phải thu thập dữ liệu giá tại mọi thời điểm. Nó không chỉ là dữ liệu lớn mà còn là “văn bản lớn”, bởi hệ thống phải phân tích các từ để nhận ra khi nào một sản phẩm bị ngưng hoặc một mẫu mới sắp được tung ra, thông tin mà người tiêu dùng nên biết, và ảnh hưởng đến giá.

Một năm sau, Decide.com đã phân tích 4 triệu sản phẩm, sử dụng hơn 25 tỷ lượt theo dõi giá. Nó phát hiện những điều kỳ quặc về bán lẻ mà mọi người đã không hề “nhìn thấy” trước đây, như thực tế là giá của những kiểu mẫu cũ có thể tăng tạm thời khi những kiểu mẫu mới được tung ra. Hầu hết mọi người muốn mua một kiểu mẫu cũ hơn vì nghĩ rằng nó rẻ hơn, nhưng tùy thuộc vào lúc họ nhấp vào “mua”, họ có thể phải trả nhiều tiền hơn. Khi các cửa hàng trực tuyến ngày càng sử dụng nhiều các hệ thống giá tự động, Decide.com có thể nhận ra được những đợt tăng giá bất thường theo thuật toán và cảnh báo người tiêu dùng nên chờ đợi. Các dự đoán của công ty, theo những tính toán nội tại, là chính xác trong 77 phần trăm các trường hợp và giúp người mua tiết kiệm trung bình khoảng $100 mỗi sản phẩm. Vì vậy, công ty tự tin tới mức trong trường hợp dự đoán của nó không đúng, Decide.com sẽ hoàn trả lại phần chênh lệch giá cho

các thành viên trả phí của dịch vụ.



Phim minh họa cơ chế hoạt động của Decide.com

Điều khiến Decide.com đặc biệt không phải là dữ liệu, vì công ty này dựa vào thông tin được phép lấy miễn phí từ các trang web thương mại điện tử cũng như thông tin lấy được từ mạng. Nó cũng không phải là trình độ kỹ thuật, vì công ty không làm bất cứ điều gì phức tạp đến nỗi chỉ các kỹ sư nào đó trên thế giới mới thực hiện nổi. Thay vào đó, bản chất của điều khiến Decide.com đặc biệt là ý tưởng: công ty này có một "tư duy dữ-liệu-lớn". Nó thấy một cơ hội và phát hiện ra rằng một số dữ liệu nhất định có thể được khai thác để tiết lộ những bí mật có giá trị. Và nếu bạn thấy dường như có nét tương đồng giữa Decide.com với trang web dự đoán giá vé máy bay Farecast, quả là có lý do: cả hai đều là những đứa con tinh thần của Oren Etzioni.

Trong chương trước, chúng ta đã lưu ý rằng dữ liệu đang trở thành một nguồn mới của giá trị, phần lớn vì những gì chúng ta gọi là giá trị tương lai, khi nó bổ sung những mục đích sử dụng mới. Trọng tâm nằm ở các công ty thu thập dữ liệu. Bây giờ chúng ta chuyển sự chú ý sang các công ty sử dụng dữ liệu, và xem họ phù hợp với chuỗi giá trị thông tin như thế nào. Chúng ta sẽ xem xét điều này có nghĩa như thế nào đối với các tổ chức và cá nhân, cả trong sự nghiệp và cuộc sống hàng ngày của họ.

Có ba loại công ty dữ-liệu-lớn đã nảy sinh, được phân biệt theo giá trị mà họ cung cấp: dữ liệu, các kỹ năng, và các ý tưởng.

Đầu tiên là dữ liệu. Đây là những công ty có dữ liệu hoặc ít nhất là có thể truy cập được nó. Nhưng có lẽ đó không phải là những gì họ kinh doanh. Họ không nhất thiết phải có những kỹ năng thích hợp để tận dụng giá trị của nó hoặc để tạo ra các ý tưởng sáng tạo về những gì đáng được giải phóng. Ví dụ tốt nhất là Twitter, rõ ràng rất hứng thú với một lượng lớn dữ liệu chạy qua các máy chủ của nó, nhưng rồi được chuyển cho hai công ty độc lập để cấp phép cho những người khác sử dụng.

Thứ hai là kỹ năng. Thông thường những công ty tư vấn, các nhà cung cấp công nghệ, và các nhà cung cấp dịch vụ phân tích là những người có chuyên môn đặc biệt, nhưng tự bản thân họ lại không có dữ liệu cũng như khả năng để đề xuất những ứng dụng sáng tạo nhất cho nó. Ví dụ trong trường hợp của Walmart và Pop-Tarts, các nhà bán lẻ đã tìm đến các chuyên gia của Teradata, một công ty phân tích dữ liệu, để giúp tìm kiếm những hiểu biết sâu sắc.

Thứ ba là tư duy dữ-liệu-lớn. Với một số doanh nghiệp nhất định, dữ liệu và các bí quyết không phải là những nguyên nhân chính cho sự thành công của họ. Điều khiến họ khác biệt là những người sáng lập và nhân viên của họ có những ý tưởng độc đáo về cách khai thác dữ liệu để mở khóa cho các loại hình mới

của giá trị. Một ví dụ là Pete Warden, người đồng sáng lập khá kỳ dị của Jetpac, một công ty tư vấn du lịch dựa trên những bức ảnh người sử dụng tải lên trang web.

Cho đến nay, hai yếu tố được chú ý nhất là kỹ năng (mà ngày nay đang khan hiếm), và dữ liệu (có vẻ rất phong phú). Một nghề chuyên môn mới đã xuất hiện trong những năm gần đây, "nhà khoa học dữ liệu", kết hợp các kỹ năng của nhà thống kê, người lập trình phần mềm, nhà thiết kế thông tin đồ họa, và người kể chuyện. Thay vì dán mắt vào kính hiển vi để mở khóa một bí ẩn của vũ trụ, nhà khoa học dữ liệu kết bạn với các cơ sở dữ liệu để tạo nên một khám phá. Học viện McKinsey Global đã đưa ra những dự đoán bi quan về sự khan hiếm các nhà khoa học dữ liệu trong cả hiện tại và tương lai (điều mà các nhà khoa học dữ liệu đang thích trích dẫn để cảm thấy mình đặc biệt và để đòi hỏi tăng lương).

Hal Varian, nhà kinh tế trưởng của Google, đã gọi nghề thống kê là công việc "gợi cảm nhất". "Nếu muốn thành công, bạn sẽ muốn mình vừa là phần bổ khuyết vừa là thứ khan hiếm trong một lĩnh vực phổ biến và không đắt đỏ", ông nói. "Dữ liệu đang có sẵn một cách rộng rãi và rất quan trọng về chiến lược đến mức thứ đang khan hiếm chính là những kiến thức để trích xuất trí tuệ từ nó. Đó là lý do các nhà thống kê, quản lý cơ sở dữ liệu và chuyên gia 'dạy máy tính học' (machine learning) sẽ thực sự có một vị trí tuyệt vời". Tuy nhiên, tất cả việc tập trung vào các kỹ năng và hạ thấp tầm quan trọng của dữ liệu có thể sẽ thất bại. Khi ngành công nghiệp lớn mạnh, sự thiếu nhân công sẽ được khắc phục khi các kỹ năng như Varian tuyên bố trở nên phổ biến. Hơn nữa, có một niềm tin sai lầm rằng chỉ vì có quá nhiều dữ liệu xung quanh, nên việc lấy dữ liệu là miễn phí hoặc giá trị của nó là ít ỏi. Trong thực tế, dữ liệu là thành phần quan trọng. Để hiểu được tại sao, hãy xem xét các thành phần khác nhau của chuỗi giá trị dữ-liệu-lớn, và chúng sẽ có thể thay đổi như thế nào theo thời gian. Để bắt đầu, chúng ta hãy xem xét lần lượt từng

loại - người sở hữu dữ liệu, chuyên gia dữ liệu, và tu duy dữ-liệu-lớn.

Chuỗi giá trị dữ-liệu-lớn

Thứ cơ bản tạo nên dữ liệu lớn là bản thân thông tin. Chủ sở hữu dữ liệu có thể không thực hiện công việc sưu tập ban đầu, nhưng họ kiểm soát việc truy cập thông tin và sử dụng nó cho chính họ hoặc cấp phép cho những người khác để tận dụng giá trị của nó. Ví dụ ITA Software, một mạng giữ chỗ hàng không lớn (sau Amadeus, Travelport, và Sabre), đã cung cấp dữ liệu cho Farecast để dự báo giá vé máy bay, nhưng không tự tiến hành các phân tích. Tại sao không? ITA nhìn nhận việc kinh doanh của họ là sử dụng dữ liệu cho các mục đích mà nó được thiết kế - bán vé máy bay - chứ không phải cho các ứng dụng phụ trợ. Như vậy, những năng lực cốt lõi của nó là khác. Hơn nữa, nó sẽ phải làm việc xung quanh bằng sáng chế của Etzioni.

Công ty cũng quyết định không khai thác dữ liệu do vị trí của nó trong chuỗi giá trị thông tin. “ITA tránh xa các dự án nào khiến cho việc ứng dụng thương mại của dữ liệu có liên quan chặt chẽ tới doanh thu trong ngành hàng không”, Carl de Marcken, người đồng sáng lập và cựu giám đốc công nghệ của ITA Software, nhớ lại. “ITA có quyền truy cập đặc biệt tới loại dữ liệu như vậy vì chúng rất cần cho việc cung cấp dịch vụ của ITA, nên không thể làm phương hại điều này”. Thay vào đó, công ty cấp phép dữ liệu nhưng không sử dụng nó. Phần lớn giá trị thứ cấp của dữ liệu là cho Farecast: cho khách hàng của công ty này, dưới hình thức vé máy bay rẻ hơn; cho các nhân viên và chủ sở hữu của Farecast từ thu nhập mà công ty kiếm được nhờ quảng cáo, các khoản hoa hồng, và cuối cùng là việc bán công ty.

Một số công ty đã khôn ngoan định vị bản thân ở trung tâm của dòng thông tin để họ có thể đạt được quy mô và nắm bắt giá trị từ dữ liệu. Đó là trường hợp của ngành công nghiệp thẻ tín dụng

ở Hoa Kỳ. Trong nhiều năm, chi phí cao trong việc chống gian lận khiến nhiều ngân hàng nhỏ và vừa tránh phát hành thẻ tín dụng của mình và giao hoạt động thẻ của mình cho các tổ chức tài chính lớn hơn, có đủ lực để đầu tư vào công nghệ. Các công ty như Capital One và Ngân hàng MBNA của Bank of America đã chiếm được thị trường này. Những ngân hàng nhỏ hơn bây giờ hối tiếc về quyết định đó, vì việc bỏ các hoạt động thẻ sẽ tước đi của họ dữ liệu về các mẫu chi tiêu có thể giúp hiểu nhiều hơn về khách hàng, từ đó bán những dịch vụ thích hợp.

Ngược lại, các ngân hàng lớn và các tổ chức phát hành thẻ như Visa và MasterCard có vẻ đã kiếm được món hời chuỗi giá trị thông tin. Bằng cách phục vụ nhiều ngân hàng và thương nhân, họ có thể nhìn thấy nhiều giao dịch hơn trên các mạng của họ và sử dụng chúng để suy luận về hành vi của người tiêu dùng. Mô hình kinh doanh của họ chuyển từ việc chỉ đơn giản xử lý thanh toán sang thu thập dữ liệu. Câu hỏi đặt ra sau đó là họ làm những gì với nó.

MasterCard có thể cấp phép dữ liệu cho các bên thứ ba để tận dụng giá trị, như ITA đã làm, nhưng công ty này thích tự phân tích. Một bộ phận được gọi là MasterCard Advisors tập hợp và phân tích 65 tỷ giao dịch từ 1,5 tỷ chủ thẻ ở 210 quốc gia để tiên đoán các xu hướng kinh doanh và tiêu dùng. Sau đó, nó bán thông tin này cho những người khác. Nó phát hiện ra, trong nhiều thứ khác, rằng nếu người ta đổ bình xăng vào khoảng 4 giờ chiều, thì họ rất có thể sẽ chi tiêu từ $35 tới $50 trong giờ kế tiếp tại một cửa hàng tạp hóa hay tiệm ăn. Một nhà tiếp thị có thể sử dụng hiểu biết này để in trên mặt sau của các hóa đơn bán xăng những tờ phiếu giảm giá cho một siêu thị gần đó, vào khoảng 4 giờ chiều.

Với tư cách một người trung gian của các dòng thông tin, MasterCard ở một vị trí tốt hơn hết để thu thập dữ liệu và nắm bắt giá trị của nó. Người ta có thể tưởng tượng một tương lai khi

các công ty thẻ bỏ các khoản hoa hồng của họ trên các giao dịch, xử lý chúng miễn phí để đổi lấy quyền truy cập vào nhiều dữ liệu hơn, và kiếm thu nhập từ việc bán các phân tích rất tinh vi dựa trên nó.

Nhóm thứ hai bao gồm các chuyên gia về dữ liệu: các công ty với chuyên môn hoặc công nghệ để thực hiện các phân tích phức tạp. MasterCard đã chọn tự thực hiện điều này trong nội bộ, còn một số công ty thì chuyển qua lại giữa các thể loại. Nhưng rất nhiều công ty khác tìm đến các chuyên gia. Ví dụ công ty tư vấn Accenture làm việc với các công ty trong nhiều ngành công nghiệp để triển khai các công nghệ cảm-biến-vô-tuyến cao cấp, và để phân tích dữ liệu các cảm biến thu thập. Trong một dự án thí điểm với thành phố St Louis, bang Missouri, Accenture cài đặt các bộ cảm biến vô tuyến trong xe buýt công cộng để giám sát động cơ nhằm dự đoán sự cố hoặc xác định thời gian tối ưu để bảo trì thường kỳ. Nó đã giúp giảm được 10 phần trăm chi phí. Chỉ một phát hiện - đó là thành phố có thể trì hoãn lịch trình thay thế phụ tùng xe từ mỗi 200.000 - 250.000 dặm lên 280.000 dặm - đã tiết kiệm được hơn 1.000 đôla cho mỗi chiếc xe. Khách hàng, chứ không phải là công ty tư vấn, đã thu hoạch được quả ngọt từ dữ liệu.

Trong lĩnh vực dữ liệu y tế, chúng ta thấy một ví dụ nổi bật khác về cách thức các công ty không thuộc lĩnh vực công nghệ có thể cung cấp các dịch vụ hữu ích như thế nào. Trung tâm MedStar Washington ở Washington DC làm việc với Microsoft Research và sử dụng phần mềm Amalga của Microsoft đã phân tích ẩn danh hồ sơ y tế trong nhiều năm - về nhân khẩu học bệnh nhân, các kiểm tra, chẩn đoán, điều trị, và nhiều thứ khác nữa - để tìm cách giảm tỷ lệ tái phát và nhiễm trùng. Đây là một số trong các khâu tốn kém nhất của chăm sóc sức khỏe, vì vậy bất cứ điều gì có thể làm giảm những tỷ lệ này đều giúp tiết kiệm rất nhiều.

Kỹ thuật này phát hiện những mối tương quan đáng ngạc nhiên.

Một trong số kết quả của nó là danh sách tất cả các điều kiện làm tăng khả năng một bệnh nhân xuất viện sẽ phải quay lại trong vòng một tháng. Có những điều người ta đã biết, nhưng chưa tìm được giải pháp dễ dàng, ví dụ một bệnh nhân suy tim sung huyết sẽ nhiều khả năng phải trở lại, do nó là một tình trạng khó điều trị. Tuy nhiên hệ thống cũng phát hiện một chỉ báo hàng đầu bất ngờ: trạng thái tinh thần của bệnh nhân.

Xác suất để một người sẽ nhập viện trở lại trong vòng một tháng sau khi xuất viện tăng lên đáng kể nếu lời khai ban đầu bao gồm những từ ngữ gợi đến suy nhược tinh thần, chẳng hạn như "chán nản". Mặc dù mối tương quan này không cho biết điều gì để thiết lập quan hệ nhân quả, tuy nhiên nó cho thấy một sự can thiệp sau khi xuất viện để giải quyết sức khỏe tâm thần của bệnh nhân sẽ có thể cải thiện được sức khỏe thể chất của họ, làm giảm tỷ lệ nhập viện trở lại và giảm chi phí y tế. Phát hiện này, mà máy tính sàng lọc ra được từ kho tàng dữ liệu khổng lồ, là điều mà một người nghiên cứu dữ liệu có thể không bao giờ phát hiện ra. Microsoft không kiểm soát dữ liệu, nó thuộc về bệnh viện. Và Microsoft đã không có một ý tưởng gì đáng kinh ngạc, vì đó không phải là những gì cần đòi hỏi ở đây. Thay vào đó, nó cung cấp công cụ, phần mềm Amalga, để phát hiện những hiểu biết sáng suốt.

Các công ty là chủ sở hữu dữ-liệu-lớn dựa vào các chuyên gia để khai thác giá trị từ dữ liệu. Nhưng bất chấp những lời khen ngợi và các chức danh sang trọng như "ninja dữ liệu", cuộc sống của các chuyên gia kỹ thuật không phải lúc nào cũng hấp dẫn như vẻ bề ngoài. Họ làm việc quần quật trong các mỏ kim cương của dữ liệu lớn, mang về nhà khoản tiền lương dễ chịu, nhưng họ sẽ trao lại những viên đá quý mà họ khai quật được cho những người có dữ liệu.

Nhóm thứ ba được tạo thành từ các công ty, cá nhân có một tư duy dữ liệu lớn. Sức mạnh của họ là ở chỗ họ nhìn thấy cơ hội

trước khi những người khác nhìn thấy - ngay cả khi thiếu các dữ liệu hoặc kỹ năng để hành động theo những cơ hội đó. Có lẽ lý do chính xác là vì với tư cách những kẻ ngoại đạo, họ thiếu những thứ giam hãm tâm trí, và được tự do tưởng tượng: họ nhìn thấy những khả năng thay vì bị giới hạn bởi những gì khả thi.

Bradford Cross thể hiện thế nào là có một tư duy dữ-liệu-lớn. Tháng 8 năm 2009, khi mới ngoài hai mươi, ông và một số bạn bè đã thành lập FlightCaster.com. Giống như FlyOnTime. us, FlightCaster dự đoán liệu một chuyến bay ở Mỹ có khả năng bị chậm trễ không. Để thực hiện các dự đoán, nó phân tích tất cả các chuyến bay trong mười năm trước, đối chiếu với dữ liệu thời tiết lịch sử và hiện tại.

Thật thú vị là các chủ sở hữu dữ liệu tự bản thân không thể làm điều đó. Không ai có động lực - hoặc nhiệm vụ bắt buộc - để sử dụng dữ liệu theo cách này. Thực tế, nếu các nguồn dữ liệu - Văn phòng Thống kê Giao thông Vận tải Mỹ, Cục Hàng không Liên bang, và Cục Thời tiết Quốc gia - dám dự báo sự chậm trễ của các chuyến bay thương mại, biết đâu Quốc hội sẽ tổ chức các phiên điều trần và các vị quan liêu sẽ xoay như chong chóng. Còn các hãng hàng không thì chẳng thể làm điều đó - hoặc không muốn làm. Họ được hưởng lợi từ việc giữ hiệu suất hoạt động bình thường của mình mơ hồ nhất có thể. Thật ra chuyện đó cũng cần cả tá kỹ sư. Trên thực tế, các dự báo của FlightCaster chính xác một cách kỳ lạ đến mức ngay cả nhân viên hãng hàng không cũng bắt đầu sử dụng chúng. Vào tháng Giêng năm 2011, Cross và các đối tác của ông đã bán công ty cho Next Jump, một công ty quản lý các chương trình giảm giá sử dụng các kỹ thuật dữ-liệu-lớn.

Khái niệm về tư duy dữ-liệu-lớn, và vai trò của một người ngoài cuộc sáng tạo với một ý tưởng tuyệt vời, không khác với những gì đã xảy ra vào buổi bình minh của thương mại điện tử trong giữa những năm 1990, khi những người tiên phong đã không bị

cản trở bởi những suy nghĩ cố thủ hoặc những ràng buộc thể chế của các ngành công nghiệp cũ. Vì vậy, một quỹ đầu tư, chứ không phải Barnes & Noble, đã thành lập một hiệu sách trực tuyến (Jeff Bezos của Amazon). Một nhà phát triển phần mềm, chứ không phải Sotheby, đã xây dựng một trang web đấu giá (Pierre Omidyar của eBay). Ngày nay, các doanh nhân với tư duy dữ-liệu-lớn thường không có dữ liệu khi họ bắt đầu. Nhưng vì thế, họ cũng không có những quyền lợi hoặc mất mát tài chính để ngăn cản việc bộc lộ những ý tưởng của họ.

Như chúng ta đã nhìn thấy, có những trường hợp trong đó một công ty kết hợp nhiều đặc điểm dữ-liệu-lớn. Có thể Etzioni và Cross đã nảy ra những ý tưởng xuất sắc trước những người khác, nhưng họ còn có các kỹ năng. Teradata và Accenture cũng được biết tới như những công ty tuyệt vời. Những người đi tiên phong về dữ liệu lớn ngày nay thường có xuất phát điểm khác nhau, và áp dụng phối hợp các kỹ năng dữ liệu của họ trong nhiều lĩnh vực. Một thế hệ mới các nhà đầu tư và các nhà doanh nghiệp đang nổi lên, đáng chú ý là các cựu thành viên của Google và cái gọi là PayPal Mafia (cựu lãnh đạo của công ty như Peter Thiel, Reid Hoffman, và Max Levchin). Họ, cùng với một số nhà khoa học máy tính trong giới hàn lâm, nằm trong số những người hỗ trợ lớn nhất của các công ty mới thành lập về dữ liệu hiện nay.

Tầm nhìn sáng tạo của các cá nhân và các công ty trong chuỗi dữ liệu lớn giúp chúng ta đánh giá lại giá trị của các công ty. Ví dụ Salesforce.com có thể không chỉ đơn giản là một nền tảng hữu ích cho các doanh nghiệp để chạy các ứng dụng của họ: nó còn xếp hạng cao về phát huy giá trị từ dữ liệu chảy trên cơ sở hạ tầng của nó. Các công ty điện thoại di động, như chúng ta đã thấy trong chương trước, thu thập một lượng khổng lồ dữ liệu nhưng thường không quan tâm khai thác giá trị của nó. Tuy nhiên họ có thể cấp phép cho những người khác khai thác giá trị mới từ nó - giống như Twitter quyết định cấp phép dữ liệu của nó cho hai công ty bên ngoài.

Cả Google và Amazon đều bao trùm các nhóm trong chuỗi giá trị dữ liệu lớn, nhưng chiến lược của họ thì khác nhau. Khi Google lần đầu tiên đặt ra việc thu thập tất cả các loại dữ liệu, họ đã nghĩ tới các ứng dụng thứ cấp. Những xe Street View, như chúng ta đã thấy, thu thập thông tin GPS không chỉ dành cho dịch vụ bản đồ mà còn để huấn luyện những chiếc xe tự lái. Ngược lại, Amazon tập trung hơn vào ứng dụng chính của dữ liệu và chỉ chạm đến các ứng dụng thứ cấp như một phần thưởng ngoài lề. Ví dụ hệ thống khuyến nghị của nó cũng dựa trên dữ liệu về các chuỗi nhấp chuột, nhưng công ty đã không sử dụng các thông tin để làm những chuyện phi thường như dự đoán tình trạng của nền kinh tế hay dịch cúm.

Mặc dù thiết bị đọc sách điện tử Kindle của Amazon có khả năng cho biết một trang nào đó đã được người sử dụng ghi chú và đánh dấu rất nhiều, nhưng công ty lại không bán thông tin này cho các tác giả và các nhà xuất bản. Các chuyên viên tiếp thị muốn biết những đoạn nào được ưa thích nhất và sử dụng hiểu biết đó để bán sách. Các tác giả muốn biết những đoạn nào trong sách của họ khiến hầu hết người đọc bỏ cuộc, và sử dụng thông tin đó để cải thiện tác phẩm. Các nhà xuất bản có thể phát hiện các chủ đề cho cuốn sách lớn tiếp theo. Tuy nhiên, Amazon dường như lại để lĩnh vực dữ liệu nằm bỏ hoang.

Nếu được khai thác một cách khôn ngoan, dữ liệu lớn có thể chuyển đổi mô hình kinh doanh của công ty và cách thức các đối tác lâu dài tương tác với nhau. Trong một trường hợp rất đáng lưu ý, một nhà sản xuất ôtô lớn của châu Âu đã thay đổi mối quan hệ thương mại với một nhà cung cấp phụ tùng bằng cách khai thác dữ liệu mà nhà sản xuất phụ tùng không có. (Vì chúng tôi biết được ví dụ này từ một trong những công ty chính phân tích dữ liệu, nên không thể tiết lộ tên cụ thể.)

Xe hơi ngày nay được trang bị các vi mạch, các bộ cảm biến, và phần mềm tải dữ liệu hiệu suất đến máy tính của các nhà sản

xuất ôtô khi xe đưa tới bảo hành. Điển hình xe hạng trung bây giờ có khoảng 40 bộ vi xử lý; tất cả các thiết bị điện tử của xe xe hơi chiếm một phần ba giá của nó. Điều này làm cho những chiếc xe giống như những người kế thừa các con tàu mà Maury gọi là những “đài quan sát nổi”. Khả năng thu thập dữ liệu về hoạt động của phụ tùng xe trên đường - và tận dụng dữ liệu này để cải thiện chúng - chứng tỏ là một lợi thế cạnh tranh lớn cho các công ty có thể có được thông tin.

Khi làm việc với một công ty phân tích bên ngoài, nhà sản xuất ôtô đã phát hiện một bộ cảm biến trong thùng nhiên liệu được chế tạo bởi một nhà cung cấp của Đức đang hoạt động rất kém, tạo ra nhiều lỗi báo động sai. Công ty có thể đưa thông tin đó cho nhà cung cấp và yêu cầu điều chỉnh. Trong thời đại kinh doanh lịch lãm, nó có thể làm được điều đó. Nhưng nhà sản xuất ôtô đã phải chi một khoản lớn cho chương trình phân tích của mình, nên họ muốn sử dụng thông tin này để thu hồi một số khoản đầu tư.

Họ cân nhắc các lựa chọn của mình. Liệu có nên bán dữ liệu? Thông tin sẽ được định giá như thế nào? Nếu nhà cung cấp né tránh, còn mình bị mắc kẹt với một phụ tùng hoạt động kém thì sao? Và họ biết rằng nếu bàn giao các thông tin thì các phụ tùng tương tự cho các đối thủ cạnh tranh của họ cũng sẽ được cải thiện. Việc đảm bảo rằng sự cải thiện sẽ chỉ có lợi cho những chiếc xe riêng của mình dường như là một bước đi sáng suốt hơn. Cuối cùng, nhà sản xuất xe hơi đã đưa ra một ý tưởng mới lạ. Họ tìm được một cách để cải thiện phụ tùng với phần mềm chỉnh sửa, nhận một bằng sáng chế về kỹ thuật, sau đó bán các bằng sáng chế cho nhà cung cấp kia - và giành được một khoản tiền lớn trong quá trình này.

Các trung gian dữ liệu mới

Ai là người giữ nhiều giá trị nhất trong chuỗi giá trị dữ-liệu-lớn?

Ngày nay, câu trả lời sẽ là những người có tư duy, có ý tưởng sáng tạo. Như chúng ta đã thấy từ thời kỳ dotcom, những ai có lợi thế của người đi tiên phong sẽ thực sự có thể phát triển thịnh vượng. Nhưng lợi thế này có thể không giữ được lâu. Khi kỷ nguyên của dữ liệu lớn tiến về phía trước, những người khác sẽ áp dụng tư duy này và lợi thế của những người đi tiên phong sẽ giảm, nói một cách tương đối.

Vậy thì có lẽ mấu chốt của giá trị là thực sự ở trong các kỹ năng? Xét cho cùng, một mỏ vàng sẽ không có giá trị gì nếu bạn không thể khai thác được vàng. Tuy nhiên, lịch sử máy tính lại cho thấy điều khác. Ngày nay nhu cầu về chuyên môn trong quản trị cơ sở dữ liệu, khoa học dữ liệu, phân tích, các thuật toán “dạy cho máy học” đều cao. Nhưng theo thời gian, khi dữ liệu lớn ngày càng trở thành một phần của cuộc sống thường nhật, khi các công cụ trở nên tốt hơn và dễ sử dụng hơn, và khi nhiều người hơn có được kinh nghiệm, thì giá trị của kỹ năng cũng sẽ giảm một cách tương đối. Tương tự như vậy, khả năng lập trình máy tính trở nên phổ biến hơn giữa những năm 1960 và 1980. Ngày nay, các công ty gia công phần mềm ra nước ngoài đã làm giảm giá trị của lập trình thậm chí nhiều hơn, những gì đã từng là mẫu mực của sự nhạy bén kỹ thuật thì hiện nay là một động cơ của phát triển cho người nghèo trên thế giới. Điều này không phải để nói rằng chuyên môn dữ liệu lớn là không quan trọng. Nhưng nó không phải là nguồn quan trọng nhất của giá trị, vì người ta có thể mang nó vào từ bên ngoài.

Hiện nay, trong giai đoạn đầu của dữ liệu lớn, những ý tưởng và kỹ năng dường như có giá trị lớn nhất. Nhưng cuối cùng hầu hết giá trị sẽ ở trong chính dữ liệu. Bởi vì chúng ta có thể làm được nhiều hơn với thông tin, và cũng bởi vì những người sở hữu dữ liệu sẽ biết đánh giá đúng hơn giá trị tiềm năng của tài sản họ sở hữu, nên họ sẽ giữ nó chặt hơn bao giờ hết, và sẽ tính mức giá cao khi những người ngoài truy cập. Quay lại phép ẩn dụ ở trên, xét cho cùng, bản thân vàng mới là quan trọng nhất.

Tuy nhiên, có một khía cạnh quan trọng đối với sự lớn mạnh lâu dài của các chủ sở hữu dữ liệu. Trong một số trường hợp, “các trung gian dữ liệu” sẽ xuất hiện để có thể thu thập dữ liệu từ nhiều nguồn, tập hợp lại, và sáng tạo với nó. Chủ sở hữu dữ liệu sẽ cho phép các trung gian thực hiện vai trò này bởi vì một số giá trị của dữ liệu chỉ có thể được thu hoạch thông qua họ.

Một ví dụ là Inrix, công ty phân tích giao thông ở bên ngoài Seattle. Nó biên dịch dữ liệu vị trí địa lý theo thời gian thực từ 100 triệu xe ở Bắc Mỹ và châu Âu. Dữ liệu đến từ những chiếc xe của BMW, Ford, Toyota, và những hãng khác, cũng như từ các đội xe thương mại như taxi và xe tải giao hàng. Nó cũng lấy dữ liệu từ điện thoại di động của những người lái xe (ở đây ứng dụng điện thoại thông minh miễn phí của Inrix có vai trò quan trọng: người dùng có được tin tức giao thông, đổi lại Inrix có được tọa độ của họ). Inrix kết hợp thông tin này với dữ liệu về các khuôn mẫu giao thông trong quá khứ, thời tiết, và những thứ khác như các sự kiện địa phương để dự đoán xem giao thông sẽ lưu chuyển như thế nào. Sản phẩm từ dây chuyền dữ liệu này được chuyển tiếp đến hệ thống định vị của xe, và được sử dụng bởi chính phủ và các đội xe thương mại.

Inrix là nhà trung gian dữ liệu độc lập tinh tế. Nó thu thập thông tin từ nhiều công ty xe hơi là đối thủ của nhau và do đó tạo ra một sản phẩm có giá trị hơn bất kỳ công ty nào trong số đó có thể đạt được riêng lẻ. Mỗi nhà sản xuất ôtô có thể có được vài triệu điểm dữ liệu từ những chiếc xe của nó lưu thông trên đường. Mặc dù nó có thể sử dụng dữ liệu này để dự báo lưu lượng giao thông, nhưng những dự báo này sẽ không chính xác hoặc hoàn chỉnh. Chất lượng dự báo được cải thiện khi số lượng dữ liệu tăng. Ngoài ra, các công ty xe hơi có thể không có các kỹ năng: năng lực của họ chủ yếu là uốn kim loại, chứ không phải là suy nghĩ về các phân phối Poisson.

Vì vậy, tất cả họ đều được khuyến khích tìm đến một bên thứ ba

để thực hiện công việc. Bên cạnh đó, mặc dù dự báo giao thông là quan trọng đối với người lái xe, nó hầu như không ảnh hưởng đến việc liệu ai đó sẽ mua hoặc không mua một chiếc xe cụ thể. Vì vậy, các đối thủ cạnh tranh không ngần ngại cùng tham gia theo cách này.

Tất nhiên, các công ty trong nhiều ngành công nghiệp đã chia sẻ thông tin trước đây. Các công ty nghiên cứu thị trường đã tổng hợp dữ liệu công nghiệp trong nhiều thập kỷ, cũng như các công ty chuyên về kiểm toán lưu thông báo chí. Đối với một số hiệp hội thương mại, đây là cốt lõi của những gì họ làm. Sự khác biệt hôm nay là dữ liệu bây giờ có vai trò như nguyên liệu thô đi vào thị trường, một tài sản độc lập với những gì nó đã nhắm trước đó để đo lường. Ví dụ: thông tin của Inrix hữu ích hơn những gì nó thể hiện ra ngoài. Phân tích giao thông của nó được sử dụng để đo sức khỏe của nền kinh tế địa phương bởi vì nó có thể cung cấp những hiểu biết về thất nghiệp, doanh số bán lẻ và các hoạt động giải trí. Khi sự phục hồi của kinh tế Mỹ bắt đầu chệch choạc trong năm 2011, những dấu hiệu của nó đã được phát hiện bằng việc phân tích giao thông bất chấp sự chối bỏ của các chính trị gia: đường sá giờ cao điểm ít đông đúc, cho thấy tỷ lệ thất nghiệp cao hơn. Ngoài ra, Inrix đã bán dữ liệu của nó cho một quỹ đầu tư sử dụng các mô hình giao thông xung quanh các cửa hàng bán lẻ lớn như một chỉ báo cho doanh số bán hàng của nó. Nhiều xe hơn trong khu vực sẽ tương quan với việc bán hàng tốt hơn.

Những trung gian khác cũng đang mọc lên bên trong chuỗi giá trị dữ-liệu-lớn. Một công ty xuất hiện sớm là Hitwise, sau đó được Experian mua lại, đã giao dịch với các nhà cung cấp dịch vụ Internet để thu thập dữ liệu kích chuột của họ. Dữ liệu này đã được bán chỉ bằng một khoản phí nhỏ cố định thay vì một tỷ lệ phần trăm của giá trị nó mang lại. Hitwise đã chiếm được phần lớn giá trị với tu cách là nhà trung gian. Một ví dụ khác là Quantcast, đo lưu lượng trực tuyến tới các trang web để giúp họ biết thêm về nhân khẩu học của những người ghé thăm và mô

hình sử dụng. Nó cung cấp một công cụ trực tuyến cho các trang web để các trang này có thể theo dõi khi người ta ghé thăm; đổi lại, Quantcast được tiếp cận dữ liệu để giúp nó cải thiện việc quảng cáo đúng khách hàng mục tiêu.

Một số tổ chức trung gian có thể không phải là những doanh nghiệp thương mại, ví dụ Viện Chi phí Chăm sóc Y tế được thành lập năm 2012 bởi một số công ty bảo hiểm lớn nhất của Mỹ. Dữ liệu kết hợp của họ lên đến 5 tỷ yêu cầu thanh toán liên quan đến 33 triệu người (ẩn danh). Việc chia sẻ các hồ sơ cho phép các công ty phát hiện các xu hướng không thể thấy được trong các bộ dữ liệu riêng lẻ nhỏ hơn của họ. Trong số những phát hiện đầu tiên là chi phí y tế Mỹ đã tăng nhanh hơn ba lần so với lạm phát trong năm 2009-2010, nhưng với những khác biệt rõ ràng ở mức chi tiết: giá phòng cấp cứu tăng 11 phần trăm trong khi giá của các cơ sở điều dưỡng thực chất lại giảm. Rõ ràng các công ty bảo hiểm y tế không bao giờ bàn giao dữ liệu quý giá của mình trừ khi cho một trung gian phi lợi nhuận. Những động cơ của một tổ chức phi lợi nhuận là đáng tin cậy hơn, và tổ chức đó có thể được thiết lập với tính minh bạch và trách nhiệm ngay từ trong tâm thức.

Sự đa dạng của các công ty dữ-liệu-lớn cho thấy giá trị của thông tin đang dịch chuyển như thế nào. Trong trường hợp của Decide.com, dữ liệu về giá cả được cung cấp bởi các trang web của đối tác trên cơ sở chia sẻ lợi nhuận. Decide.com kiếm được hoa hồng khi người ta mua hàng thông qua trang web, nhưng các công ty cung cấp dữ liệu cũng có được một phần. Điều này cho thấy một sự trưởng thành trong cách ngành công nghiệp làm việc với dữ liệu: Trước đây, ITA không nhận được hoa hồng trên các dữ liệu nó cung cấp cho Farecast, mà chỉ có một lệ phí cấp phép cơ bản. Hiện nay các nhà cung cấp dữ liệu có thể giành được những điều khoản hấp dẫn hơn. Với công ty tiếp theo mà Etzioni lập ra, người ta có thể cho rằng ông sẽ cố gắng tự cung cấp dữ liệu, vì giá trị đã di chuyển từ kỹ năng chuyên môn sang ý

tưởng và hiện đang di chuyển sang dữ liệu.

Các mô hình kinh doanh đang được thay đổi hoàn toàn khi giá trị chuyển đến những người kiểm soát dữ liệu. Hãng sản xuất xe hơi châu Âu đạt thỏa thuận sở hữu trí tuệ với nhà cung cấp, có một đội ngũ phân tích dữ liệu mạnh mẽ, nhưng phải làm việc với một nhà cung cấp công nghệ bên ngoài để khám phá những tri thức từ các dữ liệu. Công ty công nghệ được trả phí cho công việc của nó, nhưng hãng xe giữ phần lớn lợi nhuận. Tuy nhiên, đánh hơi thấy cơ hội, công ty công nghệ đã thay đổi mô hình kinh doanh của mình để chia sẻ một số rủi ro và phần thưởng với khách hàng. Nó đã thử nghiệm làm việc với mức phí thấp hơn để đổi lấy một số chia sẻ của cải mà phân tích của nó mang lại. (Đối với các nhà cung cấp phụ tùng xe hơi, có thể yên tâm khẳng định rằng trong tương lai tất cả họ sẽ muốn bổ sung các cảm biến đo lường vào sản phẩm của mình, hoặc nhấn mạnh quyền truy cập vào dữ liệu hiệu suất như một phần chuẩn của hợp đồng mua bán, để liên tục cải tiến các phụ tùng của họ.)

Với các nhà trung gian, công việc của họ sẽ phức tạp bởi vì họ phải thuyết phục các công ty về giá trị của việc chia sẻ. Ví dụ Inrix đã bắt đầu thu thập nhiều hơn chứ không chỉ thông tin vị trí địa lý. Năm 2012, công ty chạy một thử nghiệm phân tích xem ở đâu và khi nào thì hệ thống phanh tự động (ABS) được kích hoạt, để hãng xe hơi đã thiết kế hệ thống đo từ xa của nó thu thập thông tin trong thời gian thực. Ý tưởng là việc thường xuyên kích hoạt ABS trên một đoạn cụ thể trên đường có thể cho thấy rằng các điều kiện ở đó nguy hiểm, và người lái xe nên xem xét những tuyến đường thay thế. Vì vậy, với các dữ liệu này, Inrix có thể khuyến cáo không chỉ con đường ngắn nhất mà còn cả con đường an toàn nhất nữa.

Tuy nhiên, hãng sản xuất xe hơi không có kế hoạch chia sẻ dữ liệu với người khác. Thay vào đó, nó nhất quyết yêu cầu Inrix chỉ triển khai hệ thống độc quyền trong xe hơi của mình. Giá trị của

việc tung hô tính năng này được xem là lớn hơn những gì đạt được từ việc tổng hợp dữ liệu của nó với dữ liệu của các hãng khác để tăng độ chính xác tổng thể của hệ thống. Tuy nhiên Inrix tin rằng theo thời gian, tất cả các nhà sản xuất xe hơi sẽ thấy được tiện ích của việc tập hợp tất cả các dữ liệu của họ. Với tư cách một nhà trung gian dữ liệu, Inrix có một động lực mạnh mẽ để bám vào niềm lạc quan như vậy: hoạt động của nó được xây dựng hoàn toàn trên việc truy cập tới nhiều nguồn dữ liệu.

Các công ty cũng đang thử nghiệm các hình thức tổ chức khác nhau trong ngành dữ liệu lớn. Inrix không phải tình cờ đi theo mô hình kinh doanh này, giống như trường hợp của nhiều công ty mới thành lập khác, mà nó được thiết kế ngay từ đầu cho vai trò nhà trung gian. Microsoft, nơi sở hữu các bằng sáng chế lớn của công nghệ, đã thấy một công ty nhỏ, độc lập - chứ không phải là một công ty lớn - có thể được coi là trung lập hơn, và có thể mang các đối thủ công nghiệp lại cùng nhau và thu được nhiều nhất từ sở hữu trí tuệ của mình. Tương tự, Trung tâm Bệnh viện Washington MedStar sử dụng phần mềm Amalga của Microsoft để phân tích các tái nhập viện của bệnh nhân đã biết chính xác những gì nó đã làm với dữ liệu của nó: hệ thống Amalga ban đầu là phần mềm phòng cấp cứu nội bộ riêng của bệnh viện, được gọi là Azyxxi, rồi nó được bán cho Microsoft vào năm 2006 để có thể được phát triển tốt hơn.

Năm 2010 UPS bán một đơn vị phân tích dữ liệu nội bộ, gọi là UPS Logistics Technologies, cho công ty cổ phần tư nhân Thoma Bravo. Hiện đang hoạt động dưới tên Roadnet Technologies, đơn vị này tự do hơn để phân tích tuyến đường cho nhiều công ty. Roadnet thu thập dữ liệu từ nhiều khách hàng để cung cấp một dịch vụ điểm chuẩn cho toàn ngành công nghiệp được sử dụng bởi UPS và các đối thủ cạnh tranh của nó. UPS Logistics sẽ không bao giờ thuyết phục được các đối thủ của công ty mẹ của nó để bàn giao các bộ dữ liệu của họ, giám đốc điều hành Roadnet Len Kennedy giải thích. Tuy nhiên sau khi Roadnet trở thành độc

lập, các đối thủ cạnh tranh của UPS cảm thấy thoải mái hơn khi cung cấp dữ liệu của họ, và cuối cùng tất cả mọi người đều hưởng lợi từ độ chính xác được cải thiện nhờ việc tập hợp dữ liệu mang lại.

Có thể tìm thấy bằng chứng về việc bản thân dữ liệu, chứ không phải các kỹ năng hay tư duy, sẽ được định giá cao nhất trong nhiều vụ chuyển nhượng doanh nghiệp dữ-liệu-lớn. Ví dụ trong năm 2006, Microsoft tưởng thưởng tư duy dữ-liệu-lớn của Etzioni qua việc mua Farecast với giá khoảng 110 triệu USD. Nhưng hai năm sau đó Google đã chi 700 triệu USD để mua lại nhà cung cấp dữ liệu của Farecast, ITA Software.

Sự cáo chung của các chuyên gia

Trong bộ phim *Moneyball,* đội bóng chày Oakland A's đã trở thành người chiến thắng bằng cách áp dụng các phân tích và các loại số liệu mới vào. Có một cảnh thú vị trong đó các tuyển trạch viên già tóc hoa râm đang ngồi xung quanh một chiếc bàn thảo luận về các cầu thủ. Khán giả chắc hẳn rụng rời, không chỉ vì cảnh này cho thấy các quyết định được đưa ra chẳng dựa trên dữ liệu, mà còn bởi vì chúng ta đều từng rơi vào những tình huống trong đó sự "chắc chắn" chỉ dựa trên tình cảm thay vì khoa học.

"Anh ta có một cơ thể bóng chày... một khuôn mặt đẹp", một tuyển trạch viên nói.

"Anh ta có một cú vung chày tuyệt vời. Khi tiếp xúc bóng, anh ta đánh mạnh, bóng bật giòn khỏi chày", một ông tóc bạc mang máy trợ thính nói thều thào. "Rất nhiều tiếng bật giòn khỏi chày", ông kia đồng tình.

Một ông thứ ba cắt ngang cuộc hội thoại, tuyên bố: "Anh ta có bạn gái xấu òm". "Thế nghĩa là sao?", tuyển trạch viên chủ trì cuộc họp hỏi. "Bạn gái xấu nghĩa là không tự tin", người phản đối

giải thích cứ như chuyện đó thật hiển nhiên. "OK", người chủ trì hài lòng nhận xét, và tiếp tục.

Sau khi đùa cợt, một tuyển trạch viên đến nay vẫn im lặng nói: "Anh chàng này có tinh thần. Đó là điều tốt. Tôi muốn nói hắn là kiểu mà khi ta bước vào thì hắn đã có mặt ở đó sẵn rồi". Một người khác thêm vào: "Anh ta hấp dẫn đấy. Anh ta trông đẹp trai, sẵn sàng tham gia. Anh ta chỉ cần được chơi một thời gian". "Tôi chỉ muốn nói", người phản đối nhắc lại, "bạn gái của anh ta chỉ điểm sáu - tối đa!".

Cảnh này mô tả hoàn hảo những thiếu sót trong phán xét của con người. Cuộc tranh luận thực sự dựa trên những thứ chẳng hề cụ thể. Các quyết định về giá trị hàng triệu đôla của các hợp đồng mua bán cầu thủ được thực hiện theo bản năng, thiếu vắng các biện pháp khách quan. Đúng, đó chỉ là một bộ phim, nhưng thực tế cuộc sống không khác nhiều. Kiểu lý luận rỗng tuếch tương tự vẫn được sử dụng trong các phòng họp ở Manhattan hay Phòng Bầu dục, từ các quán cà phê đến các bàn ăn ở khắp mọi nơi.

Moneyball, dựa trên cuốn sách của Michael Lewis, kể về câu chuyện có thật của Billy Beane, tổng giám đốc của Oakland A's. Trong bối cảnh rối ren của đội bóng, Beane đã mang đến văn phòng quản lý của đội phương pháp quan sát thống kê (sabermetrics), thuật ngữ được nhà báo thể thao Bill James đặt ra khi nhắc đến Hiệp hội Nghiên cứu Bóng chày Mỹ, mà lúc đó vẫn bị xem như một nhóm lập dị. Beane đã thách thức giáo điều lâu đời, cũng giống như quan điểm nhật tâm của Galileo đã thách thức uy quyền của Nhà thờ Công giáo. Cuối cùng Beane đã dẫn dắt đội bóng giành vị trí số một ở giải miền Tây nước Mỹ trong mùa giải 2002, trong đó có 20 trận thắng liên tiếp. Từ đó, các nhà thống kê đã thay thế các tuyển trạch viên để trở thành các chuyên gia thể thao. Rất nhiều đội khác sau đó cũng tự áp dụng phuong pháp quan sát thống kê này.

Cũng với tinh thần đó, tác động lớn nhất của dữ liệu lớn sẽ là các quyết định dựa vào dữ liệu sẽ được đưa ra để củng cố hoặc bác bỏ phán quyết của con người.

Trong cuốn sách *Super Crunchers (Những nhà phân tích dữ liệu siêu đẳng)*, luật sư và nhà kinh tế học trường Yale Ian Ayers cho rằng phân tích thống kê buộc người ta phải xem xét lại bản năng của họ. Thông qua dữ liệu lớn, điều này càng trở nên cần thiết hơn. Các chuyên gia trong ngành sẽ mất đi hào quang của mình so với nhà thống kê và nhà phân tích dữ liệu, là những người được giải phóng khỏi những cách cũ để làm việc và để cho các dữ liệu tự nói. Những người này sẽ dựa trên các mối tương quan mà không cần các phán đoán và thành kiến, giống như Maury đã không chọn giá trị bề mặt từ những gì các thủy thủ đã nói về một tuyến đường nhất định, nhưng tin tưởng vào số liệu tổng hợp để tiết lộ những sự thật thực tế.

Chúng ta đang nhìn thấy sự suy tàn ảnh hưởng của các chuyên gia chuyên ngành trong nhiều lĩnh vực. Trong truyền thông, các nội dung được tạo ra và xuất bản trên các trang web như Huffington Post, Gawker, và Forbes thường xuyên được xác định bởi dữ liệu, chứ không chỉ bởi sự phán xét của các biên tập viên. Dữ liệu có thể tiết lộ những gì mọi người muốn đọc tốt hơn so với bản năng của các nhà báo dày dạn. Công ty đào tạo trực tuyến Coursera sử dụng thông tin về việc sinh viên xem lại phần nào trong bài giảng để tìm hiểu những nội dung nào có thể đã không rõ ràng, và phản hồi lại cho giáo viên để họ cải thiện. Như chúng ta đã thấy trước đây, Jeff Bezos loại bỏ các nhân viên điểm sách tại Amazon khi dữ liệu cho thấy các khuyến cáo theo thuật toán đã mang lại nhiều doanh thu hơn.

Điều này có nghĩa các kỹ năng cần thiết để thành công tại nơi làm việc đang thay đổi. Nó làm thay đổi những gì nhân viên được trông đợi sẽ mang đến cho các tổ chức của họ. Tiến sĩ McGregor, người chăm sóc cho trẻ sinh non ở Ontario, không cần

là bác sĩ thông thái nhất tại bệnh viện, hoặc là người có thẩm quyền cao nhất về chăm sóc trẻ sơ sinh trên thế giới, để mang lại những kết quả tốt nhất cho bệnh nhân của bà. Thật ra, bà không phải là một bác sĩ - bà có bằng tiến sĩ về khoa học máy tính. Nhưng bà áp dụng dữ liệu của hơn một thập kỷ về bệnh nhân, và máy tính đã nghiền ngẫm chúng để bà biến chúng thành các kiến nghị trong điều trị.

Như chúng ta đã thấy, những người tiên phong trong dữ liệu lớn thường đến từ các ngành nghề bên ngoài lĩnh vực mà họ làm nên tên tuổi. Họ là những chuyên gia trong phân tích dữ liệu, trí tuệ nhân tạo, toán học, hoặc thống kê, và họ áp dụng những kỹ năng này vào các ngành công nghiệp cụ thể. Những người chiến thắng của các cuộc thi Kaggle, nền tảng trực tuyến cho các dự án dữ-liệu-lớn, thường chỉ mới tiếp xúc với lĩnh vực mà trong đó họ tạo ra được những thành công, Giám đốc điều hành Anthony Goldbloom của Kaggle giải thích. Một nhà vật lý người Anh đã phát triển các thuật toán suýt giành chiến thắng, để dự đoán yêu cầu thanh toán bảo hiểm và xác định các xe hơi cũ có lỗi. Một chuyên gia bảo hiểm Singapore dẫn đầu một cuộc thi dự đoán các phản ứng sinh học đối với các hợp chất hóa học. Trong khi đó, ở nhóm dịch máy của Google, các kỹ sư vui mừng với bản dịch của những ngôn ngữ mà chẳng ai trong văn phòng nói được. Tương tự như vậy, các nhà thống kê tại nhóm dịch máy của Microsoft thích thú đưa ra một lời châm biếm cũ: rằng chất lượng của bản dịch tăng mỗi khi một nhà ngôn ngữ học rời khỏi nhóm.

Chắc chắn là các chuyên gia chuyên ngành sẽ không biến mất. Nhưng uy quyền của họ sẽ suy giảm. Từ nay, họ phải chia sẻ diễn đàn với các chuyên viên dữ-liệu-lớn, cũng giống như quan hệ nhân quả tráng lệ phải chia sẻ ánh đèn sân khấu với mối tương quan khiêm nhường. Điều này làm biến đổi cách chúng ta đánh giá kiến thức, bởi vì chúng ta có xu hướng nghĩ rằng những nhà chuyên môn sâu có giá trị cao hơn những người nghiên cứu rộng

- rằng thời vận ủng hộ chiều sâu. Tuy nhiên, chuyên môn giống như sự chính xác: thích hợp cho một thế giới dữ-liệu-nhỏ nơi ta không bao giờ có đủ thông tin, hoặc thông tin đúng, và do đó phải dựa trên trực giác và kinh nghiệm để dẫn đường. Trong một thế giới như vậy, kinh nghiệm đóng một vai trò quan trọng, vì nó là sự tích lũy lâu dài kiến thức tiềm ẩn - kiến thức mà người ta không thể dễ dàng truyền tải hoặc học hỏi từ một cuốn sách, hoặc thậm chí có ý thức về nó. Loại kiến thức đó giúp con người ra quyết định thông minh hơn.

Nhưng khi bị nhồi nhét điên khùng với dữ liệu, bạn có thể khai thác nó, và với hiệu quả lớn hơn. Vì vậy, những người phân tích dữ liệu lớn có thể nhìn xa hơn các tín điều và suy nghĩ thông thường, không phải vì họ thông minh hơn, mà vì họ có dữ liệu. (Và là những người ngoài, họ không thiên vị những luận điểm chuyên môn, vốn có thể thu hẹp tầm nhìn của một chuyên gia vào bất cứ bên nào của cuộc tranh luận.) Điều này cho thấy những tiêu chuẩn xác định giá trị của một nhân viên trong công ty sẽ thay đổi. Những gì bạn cần biết sẽ thay đổi, người mà bạn cần biết sẽ thay đổi, và những gì bạn cần học để chuẩn bị cho nghề nghiệp và cuộc sống cũng thay đổi.

Toán học và thống kê, có lẽ với một chút khoa học về lập trình và mạng, sẽ là nền tảng cho công sở hiện đại, giống như khả năng tính toán một thế kỷ trước đây và khả năng đọc viết trước đó nữa. Trong quá khứ, để thành một nhà sinh vật học xuất sắc người ta cần phải biết rất nhiều nhà sinh vật học khác. Điều đó không thay đổi hoàn toàn. Tuy nhiên, ngày nay bề rộng dữ-liệu-lớn cũng quan trọng, chứ không chỉ bề sâu kiến thức chuyên môn. Việc giải một vấn đề sinh học khó rất có thể được thực hiện thông qua sự phối hợp với một nhà vật lý thiên văn hay một nhà thiết kế dữ liệu trực quan.

Trò chơi điện tử là một trong những ngành công nghiệp mà các “trung úy dữ liệu lớn” vẫn đua chen để đứng bên cạnh các “đại

tướng chuyên ngành”, đồng thời biến đổi cả ngành công nghiệp trong quá trình này. Trò chơi điện tử là lĩnh vực kinh doanh lớn, gặt hái nhiều hơn các phòng vé Hollywood hàng năm trên toàn thế giới. Trong quá khứ, các công ty thiết kế một trò chơi, phát hành nó, và hy vọng nó sẽ trở thành nổi tiếng. Theo số liệu bán hàng, các công ty sẽ chuẩn bị một phần tiếp theo hoặc bắt đầu một dự án mới. Những quyết định về nhịp độ chơi và các yếu tố của trò chơi như nhân vật, cốt truyện, các đối tượng, và sự kiện được dựa trên sự sáng tạo của các nhà thiết kế, những người thực hiện công việc của họ với cùng mức độ nghiêm túc như Michelangelo vẽ trong Nhà thờ Sistine. Đó là nghệ thuật, không phải khoa học, một thế giới của linh cảm và bản năng, rất giống như câu chuyện của các tuyển trạch viên bóng chày trong *Moneyball.*

Nhưng thời đó đã qua. FarmVille, Frontierville, FishVille, và các trò chơi khác của Zynga đều có ở dạng trực tuyến và tương tác. Xét trên bề mặt, game trực tuyến giúp Zynga biết được dữ liệu việc sử dụng, và sửa đổi các trò chơi trên cơ sở chúng được thực sự chơi như thế nào. Vì vậy, nếu người chơi gặp khó khăn khi thăng cấp, hoặc muốn bỏ ngang tại một thời điểm nào đó bởi vì tính hành động không còn hấp dẫn, Zynga có thể phát hiện những vấn đề này trong dữ liệu và khắc phục chúng. Nhưng điều ít rõ ràng hơn là công ty có thể chỉnh các trò chơi theo những đặc điểm của từng người chơi. Không chỉ có một mà có tới hàng trăm phiên bản của FarmVille.

Các nhà phân tích dữ-liệu-lớn của Zynga nghiên cứu liệu việc bán các hàng hóa ảo có bị ảnh hưởng bởi màu sắc của chúng, hoặc bởi người chơi nhìn thấy bạn bè của họ sử dụng chúng. Ví dụ sau khi dữ liệu cho thấy người chơi FishVille mua một loại cá trong suốt nhiều hơn so với các sinh vật khác 6 lần, Zynga cung cấp nhiều loài trong suốt hơn và thu lợi khá nhiều. Trong trò chơi Mafia Wars, dữ liệu cho thấy người chơi mua vũ khí với viền vàng nhiều hơn và mua hổ gia súc toàn màu trắng.

Đây không phải là điều mà một nhà thiết kế trò chơi vùi đầu trong phòng làm việc có thể biết được, mà là do dữ liệu nói. “Chúng tôi là một công ty phân tích đội lốt một công ty trò chơi điện tử. Tất cả mọi thứ được điều hành bởi các con số”, Ken Rudin, trưởng nhóm phân tích của Zynga vào thời điểm đó, giải thích trước khi anh chuyển sang phụ trách phân tích tại Facebook. Việc khai thác dữ liệu không đảm bảo cho thành công kinh doanh nhưng cho thấy những điều khả thi.

Sự chuyển hướng sang các quyết định dựa trên dữ liệu là một bước chuyển sâu sắc. Hầu hết mọi người ra quyết định dựa trên sự kết hợp của các sự kiện và suy nghĩ, cộng với khá nhiều phỏng đoán. Các nhà điều hành chỉ cần cảm thấy tự tin về quyết định của mình dựa trên bản năng thì họ đã tiến tới luôn. Nhưng điều này đang bắt đầu thay đổi khi các quyết định quản lý được thực hiện hoặc ít nhất là được xác nhận bởi mô hình dự báo và phân tích dữ-liệu-lớn.

Ví dụ The-Numbers.com sử dụng toán học và rất nhiều dữ liệu để nói với các nhà sản xuất độc lập ở Hollywood biết một bộ phim có khả năng kiếm được thu nhập bao nhiêu, từ rất sớm trước khi những cảnh đầu tiên được quay. Cơ sở dữ liệu của công ty phân tích khoảng 30 triệu hồ sơ chứa đựng tất cả các phim thương mại của Mỹ hàng thập kỷ trở lại. Dữ liệu này bao gồm ngân sách, thể loại, diễn viên, đội làm phim, và các giải thưởng, cũng như doanh thu của mỗi bộ phim (từ các phòng vé ở Mỹ và các nước, bản quyền ở nước ngoài, doanh số bán và cho thuê phim...), và nhiều nữa. Cơ sở dữ liệu cũng có một phần kết nối con người, chẳng hạn như “nhà viết kịch bản này đã làm việc với đạo diễn này, đạo diễn này đã làm việc với diễn viên kia”, Bruce Nash, người sáng lập và chủ tịch của công ty, giải thích.

The-Numbers.com có thể tìm thấy mối tương quan phức tạp dự đoán thu nhập của các dự án phim. Các nhà sản xuất đưa thông tin đó tới các hãng phim hoặc các nhà đầu tư để có được sự ủng

hộ tài chính. Công ty có thể thậm chí thao tác với các biến để nói với khách hàng làm thế nào để tăng lợi nhuận của họ (hoặc giảm thiểu rủi ro thua lỗ). Trong một trường hợp, phân tích của công ty phát hiện ra rằng một dự án sẽ có cơ hội tốt hơn nhiều để thành công nếu vai nam chính là một diễn viên hạng A: cụ thể, một diễn viên được đề cử giải Oscar được trả thù lao khoảng $5 triệu. Trong trường hợp khác, Nash đã thông báo cho IMAX rằng một phim tài liệu có thể có lãi chỉ khi ngân sách $12 triệu của nó giảm xuống thành $8 triệu. "Nó làm cho nhà sản xuất hài lòng - giám đốc thì ít hài lòng hơn", Nash nói.

Sự thay đổi trong quá trình ra quyết định của các công ty đang bắt đầu diễn ra khá rõ. Giáo sư kinh doanh tại Trường Quản trị Sloan của MIT Erik Brynjolfsson và các đồng nghiệp đã nghiên cứu hiệu suất của các công ty vượt trội về ra quyết định dựa trên dữ liệu và so sánh nó với hiệu suất của các công ty khác. Họ phát hiện ra mức năng suất cao hơn đến 6 phần trăm tại các công ty như vậy so với tại các công ty không chú trọng sử dụng dữ liệu để ra quyết định. Điều này giúp các công ty dựa trên dữ liệu có một lợi thế đáng kể - mặc dù cũng giống như lợi thế về tư duy và kỹ năng, nó sẽ không thể tồn tại lâu, vì nhiều công ty hơn sẽ áp dụng các phương pháp tiếp cận dữ-liệu-lớn cho công việc kinh doanh của họ.

Vấn đề về sự tiện ích

Khi dữ liệu lớn trở thành một nguồn lợi thế cạnh tranh cho nhiều công ty, cấu trúc của toàn bộ ngành công nghiệp sẽ được định hình lại. Tuy nhiên các phần thưởng sẽ tích lũy không đồng đều. Và những kẻ chiến thắng sẽ ở trong số các công ty lớn và nhỏ, dồn ép số đông còn lại ở giữa. Các công ty lớn như Amazon và Google sẽ tiếp tục mạnh lên. Tuy nhiên không giống như tình trạng trong thời đại công nghiệp, lợi thế cạnh tranh của họ sẽ không dựa trên quy mô vật lý. Cơ sở hạ tầng kỹ thuật to lớn của các trung tâm dữ liệu mà họ điều khiển tuy đóng vai trò quan

trọng, nhưng không phải là chất lượng quan trọng nhất của họ. Với khả năng lưu trữ kỹ thuật số phong phú và nguồn lực xử lý sẵn có để thuê với giá rẻ, có thể bổ sung chỉ trong ít phút, các công ty có thể dễ dàng điều chỉnh năng lực tính toán và lưu trữ của họ để phù hợp với nhu cầu thực tế. Việc chuyển những gì từng là chi phí cố định thành chi phí thay đổi đã làm xói mòn những lợi thế của quy mô dựa trên cơ sở hạ tầng kỹ thuật mà các công ty lớn từ lâu đã được hưởng.

Quy mô vẫn còn quan trọng, nhưng nó đã thay đổi. Điều quan trọng là quy mô của dữ liệu. Vì vậy, những chủ sở hữu dữ liệu lớn sẽ phát triển mạnh khi họ thu thập và lưu trữ nhiều hơn các nguyên liệu thô của doanh nghiệp, mà họ có thể tái sử dụng để tạo ra giá trị gia tăng.

Thách thức đối với những kẻ chiến thắng của thế giới dữ-liệu-nhỏ và với các nhà vô địch truyền thống - những công ty như Walmart, Proctor & Gamble, General Electric, Nestlé, và Boeing - là việc đánh giá cao sức mạnh của dữ liệu lớn, thu thập và sử dụng dữ liệu mang tính chiến lược hơn. Nhà sản xuất động cơ máy bay Rolls-Royce hoàn toàn thay đổi việc kinh doanh của mình trong thập kỷ qua bằng cách phân tích dữ liệu từ các sản phẩm của mình, chứ không chỉ chế tạo chúng. Từ trung tâm điều hành ở Anh, công ty liên tục giám sát hiệu suất của hơn 3.700 động cơ phản lực trên toàn thế giới để phát hiện các vấn đề trước khi sự cố xảy ra. Nó sử dụng dữ liệu để giúp biến một doanh nghiệp sản xuất thành một doanh nghiệp hai mặt: Rolls-Royce bán động cơ nhưng cũng cung cấp dịch vụ theo dõi hoạt động của chúng, tính phí cho khách hàng dựa trên thời gian sử dụng (và sửa chữa hoặc thay thế chúng trong trường hợp có vấn đề). Các dịch vụ hiện tại chiếm khoảng 70 phần trăm doanh thu hàng năm của bộ phận động cơ máy bay dân sự.

Các công ty mới thành lập cũng như các công ty đã vững mạnh, khi tham gia các lĩnh vực kinh doanh mới đều cố đặt mình vào vị

trí có thể nắm bắt những nguồn dữ liệu khổng lồ. Việc Apple thâm nhập vào ngành điện thoại di động là một ví dụ. Trước iPhone, các nhà khai thác điện thoại di động tích lũy dữ liệu sử dụng có giá trị tiềm năng từ các thuê bao nhưng không thành công trong việc tận dụng nó. Apple, ngược lại, yêu cầu trong hợp đồng với các nhà khai thác là nó sẽ phải nhận được nhiều thông tin hữu ích nhất. Bằng cách lấy dữ liệu từ các điểm của các nhà khai thác trên toàn thế giới, Apple có một bức tranh về sử dụng điện thoại di động phong phú hơn bất kỳ nhà cung cấp điện thoại di động nào khác có thể tự có được.

Dữ liệu lớn cũng mang đến những cơ hội thú vị ở đầu kia của phổ kích thước. Các đối tác nhỏ thông minh và nhanh nhẹn có thể tận hưởng "quy mô không có khối lượng", cụm từ nổi tiếng của giáo sư Brynjolfsson. Nó nghĩa là họ có thể có một sự hiện diện ảo lớn mà không cần những tài nguyên vật lý quá đắt, và có thể lan tỏa sự đổi mới một cách rộng rãi với chi phí thấp. Điều quan trọng là vì một số dịch vụ dữ-liệu-lớn tốt nhất dựa chủ yếu trên các ý tưởng sáng tạo, chúng có thể không đòi hỏi đầu tư ban đầu lớn. Các công ty nhỏ có thể cấp phép cho các dữ liệu chứ không sở hữu riêng nó, thực hiện phân tích của họ trên nền tảng điện toán đám mây không tốn kém, và nộp lệ phí cấp giấy phép với một tỷ lệ phần trăm của thu nhập kiếm được.

Nhiều khả năng các lợi ích ở cả hai đầu của phổ quy mô công ty sẽ không bị giới hạn cho người sử dụng dữ liệu mà cũng tích lũy cho các chủ dữ liệu. Những chủ sở hữu dữ liệu với quy mô lớn có động lực cao để bổ sung vào các lưu trữ dữ liệu của họ, vì làm như vậy sẽ đem lại lợi ích lớn hơn với chi phí không đáng kể. Thứ nhất, họ đã có cơ sở hạ tầng tại chỗ, để lưu trữ và xử lý. Thứ hai, có một giá trị đặc biệt trong việc kết hợp các bộ dữ liệu. Và thứ ba, một "cửa hàng bách hóa" để nhận được dữ liệu sẽ khiến mọi chuyện đơn giản hơn với những người sử dụng.

Tuy nhiên, thú vị hơn, một loại mới của các chủ dữ liệu cũng có

thể xuất hiện ở một thái cực khác: các cá nhân. Khi giá trị của dữ liệu ngày càng trở nên rõ ràng, mọi người có thể muốn thể hiện sức mạnh của họ như các chủ sở hữu thông tin gắn liền với họ - ví dụ những sở thích mua sắm của họ, những thói quen xem chương trình truyền thông, và có lẽ cả dữ liệu sức khỏe nữa. Việc sở hữu dữ liệu cá nhân có thể khiến người tiêu dùng riêng lẻ có nhiều quyền lực theo những cách chưa hề được xét tới trước đây. Người ta có thể muốn quyết định cấp phép dữ liệu của họ cho ai, và với giá bao nhiêu. Tất nhiên, không phải tất cả mọi người đều muốn cược với kẻ trả giá cao nhất; nhiều người sẽ hài lòng thấy nó được tái sử dụng miễn phí để đổi lấy dịch vụ tốt hơn, ví dụ như những lời giới thiệu chính xác về các cuốn sách trên Amazon và một kinh nghiệm sử dụng tốt hơn trên Pinterest, một dịch vụ tiếp thị sách kỹ thuật số và chia sẻ nội dung. Nhưng đối với một số lượng đáng kể những người tiêu dùng kỹ thuật số hiểu biết, ý tưởng về tiếp thị và bán thông tin cá nhân của họ có thể trở thành việc tự nhiên như viết blog, tweeting, hoặc chỉnh sửa một mục Wikipedia.

Tuy nhiên, để làm được việc này sẽ cần không chỉ một sự thay đổi sở thích của người tiêu dùng. Ngày nay, việc cấp phép dữ liệu cá nhân và cho các công ty giao dịch với mỗi cá nhân để có được dữ liệu này sẽ quá phức tạp và tốn kém. Tình huống khả thi hơn là chúng ta sẽ chứng kiến sự ra đời của các công ty mới, tập hợp dữ liệu từ nhiều người tiêu dùng, cung cấp một cách thức dễ dàng để đăng ký nó, và tự động hóa các giao dịch. Nếu chi phí của họ đủ thấp, và nếu có đủ người tin tưởng họ, ta có thể tin rằng một thị trường cho dữ liệu cá nhân sẽ được thiết lập. Những doanh nghiệp như Mydex ở Anh và các nhóm như ID3 (đồng sáng lập bởi Sandy Pentland, một đại thụ về phân tích dữ liệu cá nhân tại MIT) vẫn đang nỗ lực để biến tầm nhìn này thành hiện thực.

Cho đến khi những nhà trung gian chính thức hoạt động và người sử dụng dữ liệu bắt đầu sử dụng họ, thì những ai mong

muốn trở thành chủ sở hữu dữ liệu của riêng họ chỉ có rất ít lựa chọn. Tạm thời, để giữ lại giá trị tương lai của họ trong thời gian chờ cơ sở hạ tầng và những người trung gian hình thành, các cá nhân nên tiết lộ ít hơn chứ không phải nhiều hơn.

Tuy nhiên với các công ty cỡ vừa, dữ liệu lớn ít hữu ích hơn. Có những lợi thế về quy mô đối với công ty rất lớn, hoặc những lợi thế về chi phí và đổi mới đối với công ty nhỏ, Philip Evans của Nhóm Tư vấn Boston nhận định. Trong những lĩnh vực truyền thống, các doanh nghiệp cỡ vừa tồn tại bởi vì họ kết hợp một số kích thước tối thiểu để gặt hái những lợi ích của quy mô, cùng với một sự linh hoạt nhất định mà những công ty lớn không có. Nhưng trong một thế giới dữ-liệu-lớn, không có quy mô tối thiểu mà một công ty phải đạt được để trả cho các khoản đầu tư trong cơ sở hạ tầng sản xuất. Những công ty dữ-liệu-lớn vừa muốn giữ sự linh hoạt vừa muốn thành công sẽ thấy rằng họ không còn phải đạt được một ngưỡng về kích thước nữa. Thay vào đó, họ có thể nhỏ nhưng vẫn phát triển mạnh (hoặc được một công ty dữ-liệu-lớn khổng lồ mua lại).

Dữ liệu lớn sẽ siết chặt khúc giữa của một ngành công nghiệp, thúc đẩy các doanh nghiệp này phải trở nên rất lớn, hoặc nhỏ và nhanh, hoặc chết. Nhiều ngành nghề truyền thống cuối cùng sẽ được tái cấu trúc thành những ngành nghề dữ- liệu-lớn, từ các dịch vụ tài chính tới dược phẩm và chế tạo. Dữ liệu lớn sẽ không loại bỏ tất cả các doanh nghiệp cỡ vừa trong tất cả các lĩnh vực, nhưng nó chắc chắn sẽ gây áp lực lên các công ty trong các ngành công nghiệp có nguy cơ bị lung lay bởi sức mạnh của dữ liệu lớn.

Dữ liệu lớn cũng sẵn sàng phá vỡ những lợi thế cạnh tranh của các quốc gia. Vào thời điểm khi việc sản xuất đã bị mất về tay các nước đang phát triển, còn sự đổi mới có vẻ cũng bị tước đoạt, thì các nước công nghiệp vẫn duy trì một lợi thế vì họ nắm giữ dữ liệu và biết cách sử dụng nó. Tin xấu là lợi thế này không bền

vững. Như đã xảy ra với máy tính và Internet, bước tiên phong của phương Tây trong dữ liệu lớn sẽ giảm khi các phần khác của thế giới chấp nhận công nghệ. Tuy nhiên tin tốt cho các siêu công ty ngày nay ở các quốc gia phát triển là dữ liệu lớn sẽ có thể khuếch đại các điểm mạnh cũng như các điểm yếu của các công ty. Vì vậy, nếu một công ty làm chủ được dữ liệu lớn, nó có cơ hội không chỉ hoạt động tốt hơn các đối thủ mà còn bỏ xa họ trong vai trò dẫn đầu.

Cuộc đua vẫn còn tiếp diễn. Giống như thuật toán tìm kiếm của Google cần dữ liệu xả của người sử dụng để vận hành hiệu quả, và giống như nhà cung cấp phụ tùng xe hơi Đức đã nhìn thấy tầm quan trọng của dữ liệu để cải thiện các phụ tùng của mình, tất cả các công ty cũng có thể hưởng lợi bằng cách khai thác dữ liệu theo những cách thức thông minh.

Tuy nhiên, dù có những viễn cảnh được tô hồng, vẫn còn nhiều lý do để lo lắng. Khi dữ liệu lớn đưa ra những dự đoán ngày càng chính xác về thế giới và vị trí của chúng ta trong đó, chúng ta có thể chưa sẵn sàng cho tác động của nó đối với sự riêng tư và ý thức của chúng ta về tự do. Nhận thức và thể chế của chúng ta đã được xây dựng cho một thế giới của sự khan hiếm chứ không phải cho sự thừa thãi thông tin. Chúng ta sẽ khám phá mặt tối của dữ liệu lớn trong chương kế tiếp.

8. NHỮNG RỦI RO

TRONG GẦN BỐN MƯƠI NĂM, cho đến khi bức tường Berlin sụp đổ vào năm 1989, cơ quan an ninh quốc gia Đông Đức, được gọi là Stasi, đã do thám hàng triệu người. Sử dụng khoảng 100.000 nhân viên toàn thời gian, Stasi theo dõi từng xe hơi và đường phố. Họ mở thư và xem các tài khoản ngân hàng, nghe trộm các căn hộ và đường điện thoại. Họ xúi giục những cặp tình nhân, vợ chồng, cha mẹ và con cái giám sát lẫn nhau, phản bội niềm tin cơ bản nhất mà con người có với nhau. Các tập tin kết quả - trong đó có ít nhất 39 triệu thẻ chỉ mục và 70 dặm của các văn bản - ghi lại và trình bày chi tiết các khía cạnh mật thiết nhất trong cuộc sống của những con người bình thường. Đông Đức là một trong những quốc gia giám sát toàn diện nhất từng thấy.

Hai mươi năm sau sự sụp đổ của Đông Đức, có nhiều dữ liệu được thu thập và lưu trữ về mỗi người chúng ta hơn bao giờ hết. Chúng ta bị giám sát liên tục: khi sử dụng thẻ tín dụng để trả tiền, điện thoại di động để liên lạc, hoặc số an sinh xã hội để đăng ký. Năm 2007, giới truyền thông Anh khoái trá với sự trớ trêu rằng đã có hơn 30 camera giám sát trong phạm vi 200m của căn hộ tại Luân Đôn nơi George Orwell viết cuốn *1984*. Từ lâu trước khi Internet ra đời, các công ty chuyên ngành như Equifax, Experian, và Acxiom đã thu thập, lập bảng, và cung cấp quyền truy cập vào thông tin cá nhân cho hàng trăm triệu người trên toàn thế giới. Internet đã khiến việc theo dõi trở nên dễ dàng hơn, rẻ hơn, và hữu ích hơn. Và không chỉ có những cơ quan mờ ám của chính phủ do thám chúng ta. Amazon giám sát sở thích mua sắm, Google giám sát thói quen duyệt web, trong khi Twitter biết những gì trong tâm trí của chúng ta. Facebook dường như cũng nắm bắt tất cả những thông tin này, cùng với các mối quan hệ xã hội của người sử dụng. Các nhà khai thác điện thoại di động không chỉ biết chúng ta nói chuyện với ai, mà cả hàng xóm của họ.

Với triển vọng rằng dữ liệu lớn sẽ mang đến nhiều hiểu biết có giá trị cho những ai phân tích nó, tất cả các dấu hiệu dường như đều cho thấy mức độ những người khác thu thập, lưu trữ, và tái sử dụng dữ liệu cá nhân của chúng ta sẽ còn tăng hơn nữa. Kích thước và quy mô của các bộ sưu tập dữ liệu sẽ tăng vọt khi chi phí lưu trữ tiếp tục giảm mạnh và các công cụ phân tích trở nên mạnh mẽ hơn bao giờ hết. Nếu thời đại Internet đã đe dọa sự riêng tư thì phải chăng dữ liệu lớn sẽ khiến nó nguy hiểm nhiều hơn nữa? Đó có phải là mặt tối của dữ liệu lớn?

Đúng, và nó không phải thứ duy nhất. Ở đây, điểm quan trọng về dữ liệu lớn là một sự thay đổi về quy mô dẫn đến một sự thay đổi của trạng thái. Như chúng tôi sẽ giải thích, sự biến đổi này không chỉ khiến việc bảo vệ sự riêng tư khó khăn hơn, mà còn mang lại một mối đe dọa hoàn toàn mới: hình phạt dựa trên khuynh hướng. Đó là khả năng sử dụng dự đoán dữ-liệu-lớn về con người để phán xét và trừng phạt họ ngay cả trước khi họ hành động. Làm như vậy sẽ phủ nhận ý tưởng về sự công bằng, công lý và tự do.

Ngoài sự riêng tư và khuynh hướng, có một mối nguy hiểm thứ ba. Chúng ta có nguy cơ trở thành nạn nhân của một chế độ độc tài dữ liệu, trong đó chúng ta tôn sùng thông tin, kết quả các phân tích của chúng ta, và cuối cùng lạm dụng nó. Nếu được sử dụng một cách có trách nhiệm, dữ liệu lớn là một công cụ hữu ích trong việc ra quyết định hợp lý. Khi bị nắm giữ một cách sai trái, nó có thể trở thành một công cụ của kẻ mạnh, kẻ có thể biến nó thành một nguồn của sự đàn áp, bằng cách đơn giản là khiến khách hàng và nhân viên bực bội, hay tệ hơn là làm tổn hại đến người dân.

Các nguy cơ mất kiểm soát dữ liệu lớn liên quan đến việc tôn trọng sự riêng tư và dự báo, hoặc bị lường gạt về ý nghĩa của dữ liệu, vượt xa những chuyện vặt vãnh như các quảng cáo trực tuyến. Thế kỷ XX có một lịch sử đẫm máu với các tình huống

trong đó dữ liệu đã tiếp tay cho những kết cục bi thảm. Năm 1943 Cục Thống Kê Dân Số Hoa Kỳ bàn giao các địa chỉ khu phố (nhưng không có tên đường và số nhà, để giữ cái luận điểm bịa đặt về bảo vệ sự riêng tư) của những người Mỹ gốc Nhật nhằm dễ bắt giữ họ hơn. Những hồ sơ dân sự toàn diện nổi tiếng của Hà Lan đã được Đức quốc xã xâm lược sử dụng để bắt những người Do Thái. Những con số có năm chữ số xăm vào cánh tay của tù nhân ở trại tập trung Đức Quốc xã ban đầu là tương ứng với số thẻ đục lỗ IBM Hollerith. Việc xử lý dữ liệu đã tạo điều kiện cho tội ác giết người trên một quy mô công nghiệp.

Mặc cho sức mạnh thông tin của nó, có nhiều điều mà Stasi không thể làm được. Họ không thể biết mọi người di chuyển tới đâu ở mọi thời điểm hoặc họ đã nói chuyện với ai nếu như không có những nỗ lực rất lớn. Ngày nay, phần lớn những thông tin này được thu thập bởi các hãng điện thoại di động. Nhà nước Đông Đức, cũng như chúng ta, không thể dự đoán những ai sẽ bất đồng về chính kiến- nhưng lực lượng cảnh sát đang bắt đầu sử dụng những mô hình thuật toán để quyết định ở đâu và lúc nào thì tuần tra. Điều này cho thấy dấu hiệu về những gì sắp đến. Những xu hướng này khiến các rủi ro vốn có trong dữ liệu lớn cũng phình to như chính các bộ dữ liệu.

Làm tê liệt sự riêng tư

Rất dễ ngoại suy mối nguy hại đến tính riêng tư từ mức tăng trưởng trong dữ liệu kỹ thuật số và thấy sự tương tự với địa ngục bị giám sát của Orwell trong tác phẩm *1984*. Nhưng tình hình phức tạp hơn thế. Trước tiên, không phải tất cả dữ liệu lớn đều chứa thông tin cá nhân. Dữ liệu cảm biến từ những nhà máy lọc dầu, cũng như dữ liệu về máy móc từ sàn các nhà máy, dữ liệu về các vụ nổ hố ga hay về thời tiết sân bay không chứa những thông tin như vậy. BP và Con Edison không cần (hoặc muốn) thông tin cá nhân để đạt được giá trị từ các phân tích mà họ thực hiện. Phân tích dữ-liệu-lớn của những loại thông tin này thực tế

không đặt ra rủi ro cho sự riêng tư.

Tuy nhiên, phần lớn các dữ liệu hiện giờ được tạo ra là có bao gồm thông tin cá nhân. Và các công ty có khá nhiều động lực để thu thập nhiều hơn, giữ nó lâu hơn, và tái sử dụng nó thường xuyên. Dữ liệu có thể thậm chí không rõ ràng giống như là thông tin cá nhân, nhưng với những quá trình dữ-liệu-lớn, nó có thể dễ dàng được truy trở lại về cá nhân mà nó đề cập đến. Hoặc những chi tiết riêng tư về đời sống của một người có thể được rút ra.

Ví dụ các công ty dịch vụ tiện ích đang tung ra những “đồng hồ điện thông minh” ở Hoa Kỳ và Châu Âu để thu thập dữ liệu suốt ngày, có lẽ với tần suất mỗi sáu giây - nhiều hơn so với dòng chảy nhỏ giọt thông tin về việc sử dụng năng lượng tổng thể mà những đồng hồ truyền thống thu thập. Điều quan trọng là cách các thiết bị điện tiêu thụ năng lượng tạo ra một “chìa khóa tải” duy nhất cho thiết bị đó. Thế nên một máy đun nước nóng sẽ khác với một máy tính, và khác với đèn nuôi cần sa. Vì vậy, sự sử dụng năng lượng của một hộ gia đình sẽ tiết lộ thông tin cá nhân, có thể là cả các hành vi hàng ngày, điều kiện sức khỏe hoặc các hoạt động bất hợp pháp của cư dân. Tuy nhiên câu hỏi quan trọng không nằm ở chỗ dữ liệu lớn có làm tăng rủi ro đối với sự riêng tư hay không (câu trả lời là có), mà là liệu nó có làm thay đổi tính chất của rủi ro. Nếu mối đe dọa chỉ đơn giản là lớn hơn thì các đạo luật và quy định bảo vệ sự riêng tư có thể vẫn hiệu quả trong thời đại dữ-liệu-lớn, tất cả những gì cần làm là tăng gấp đôi nỗ lực hiện tại của chúng ta. Ngược lại, nếu vấn đề thay đổi, chúng ta có thể phải cần những giải pháp mới.

Thật không may, vấn đề đã bị biến đổi. Với dữ liệu lớn, giá trị của thông tin không còn ở duy nhất trong mục đích chính của nó. Như chúng ta đã tranh luận, hiện nay nó ở trong những ứng dụng thứ cấp. Sự thay đổi này sẽ làm suy yếu vai trò trung tâm của cá nhân trong luật bảo vệ quyền riêng tư hiện tại. Hiện nay vào thời điểm thu thập thông tin, họ được cho biết thông tin nào

sẽ được thu thập và cho mục đích gì, sau đó họ được hỏi có đồng ý hay không, để việc thu thập có thể bắt đầu. Mặc dù khái niệm "thông báo và đồng ý" này không phải là cách hợp pháp duy nhất để thu thập và xử lý dữ liệu cá nhân, theo Fred Cate, một chuyên gia về quyền riêng tư tại Đại học Indiana, nó đã biến thành một nền tảng cho các nguyên tắc bảo mật trên khắp thế giới. (Trong thực tế, nó đã tạo ra những thông báo bảo mật siêu dài mà người ta hiếm khi đọc, chứ chưa nói đến hiểu - nhưng đó lại là một câu chuyện khác).

Đáng chú ý là trong thời đại dữ-liệu-lớn, những ứng dụng thứ cấp sáng tạo nhất đã không được hình dung ra khi dữ liệu được thu thập lúc ban đầu. Làm sao các công ty có thể thông báo về một mục đích còn chưa xuất hiện? Làm sao các cá nhân có thể đồng ý về một điều chưa biết? Tuy nhiên, nếu không có sự đồng ý, bất kỳ phân tích dữ-liệu-lớn nào chứa thông tin cá nhân có thể đều phải quay trở lại từng người và xin phép cho mỗi lần tái sử dụng. Bạn thử hình dung Google sẽ cố gắng liên lạc với hàng trăm triệu người dùng để họ đồng ý cho phép sử dụng các truy vấn tìm kiếm cũ của họ cho việc dự báo dịch cúm? Không có công ty nào chịu gánh vác chi phí như vậy, ngay cả khi công việc về mặt kỹ thuật là khả thi.

Phương án thay thế là yêu cầu người dùng đồng ý với bất kỳ ứng dụng nào trong tương lai đối với dữ liệu của họ tại thời điểm thu thập, nhưng đây cũng chẳng phải cách hay. Một sự cho phép bán buôn như vậy khiến việc xin phép mất đi ý nghĩa.

Những cách thức khác để bảo vệ sự riêng tư cũng không thành. Nếu thông tin của mọi người ở trong một bộ dữ liệu thì thậm chí lựa chọn "không tham gia" có thể vẫn để lại một dấu vết. Ví dụ Street View của Google. Những chiếc xe này chụp ảnh đường phố và nhà cửa ở nhiều nước. Ở Đức, Google phải đối mặt với những phản đối rộng rãi của công chúng và phương tiện truyền thông. Người ta sợ rằng những hình ảnh nhà cửa, vườn tược của họ có

thể giúp các băng nhóm trộm cắp lựa chọn những mục tiêu hấp dẫn. Dưới áp lực luật pháp, Google đã phải đồng ý để các chủ nhà không tham gia bằng cách làm mờ nhà của họ trong ảnh. Nhưng vết xóa vẫn có thể được nhìn thấy trên Street View - bạn nhận thấy những ngôi nhà bị làm mờ đi - và kẻ trộm có thể diễn giải đó là dấu hiệu cho thấy chúng là những mục tiêu đặc biệt tốt!



Phim minh họa cơ chế vận hành của Street View

Một cách tiếp cận mang tính kỹ thuật để bảo vệ sự riêng tư - vô danh hóa - cũng không hiệu quả trong nhiều trường hợp. Vô danh hóa đề cập đến việc bỏ đi khỏi các bộ dữ liệu mọi nhận dạng cá nhân, như tên, địa chỉ, số thẻ tín dụng, ngày sinh, hoặc số an sinh xã hội. Điều này chỉ có tác dụng trong một thế giới của dữ liệu nhỏ. Chúng ta thử xem xét những trường hợp về tìm kiếm web và xếp hạng phim dường như không xác định được.

Tháng 8 năm 2006, AOL công khai phát hành một tập hợp rất lớn những truy vấn tìm kiếm cũ, với thiện chí là các nhà nghiên cứu có thể phân tích nó để có được những hiểu biết thú vị. Bộ dữ liệu gồm 20 triệu truy vấn tìm kiếm của 657.000 người sử dụng từ ngày 1 tháng 3 tới 31 tháng 5 của năm đó. Thông tin cá nhân như tên người sử dụng và địa chỉ IP đã được xóa và thay thế bằng những số định danh duy nhất. Ý tưởng là các nhà nghiên cứu có thể liên kết những truy vấn tìm kiếm của cùng một người lại với nhau, nhưng không có thông tin nhận dạng.

Tuy nhiên, trong vòng vài ngày, tờ *New York Times* đã chắp nối những lệnh tìm kiếm như “đàn ông độc thân 60” với “trà tốt cho sức khỏe” và “những người làm vườn ở Lilburn, Ga” để xác định

thành công người mang số 4417749 là Thelma Arnold, một góa phụ 62 tuổi ở Lilburn, Georgia. “Chúa ơi, đó là toàn bộ cuộc sống cá nhân của tôi”, bà nói với phóng viên *New York Times* khi ông đến gõ cửa. “Tôi không hề biết ai đó đã theo dõi mình”. Phản đối của công chúng sau đó đã dẫn đến việc sa thải Giám đốc công nghệ và hai nhân viên khác của AOL.

Tuy nhiên, chỉ hai tháng sau đó, vào tháng 10 năm 2006, dịch vụ cho thuê phim Netflix đã làm điều tương tự với sự ra mắt “giải thưởng Netflix” của họ. Công ty này đã phát hành 100 triệu hồ sơ thuê phim từ gần nửa triệu người sử dụng - và treo tiền thưởng 1 triệu USD cho bất kỳ nhóm nào có thể cải thiện hệ thống giới thiệu phim của Netflix để tăng ít nhất 10 phần trăm hiệu quả. Một lần nữa, danh tính cá nhân vẫn được lấy ra khỏi các dữ liệu. Và một lần nữa, có người vẫn bị chỉ đích danh: một người mẹ, một phụ nữ đồng tính ở vùng Trung Tây bảo thủ của Mỹ, sau đó đã kiện Netflix vì việc này dưới bí danh “Jane Doe”.

Các nhà nghiên cứu tại Đại học Texas ở Austin đã so sánh dữ liệu Netflix với những thông tin công cộng khác. Họ nhanh chóng phát hiện ra rằng những đánh giá bởi một người dùng ẩn danh trùng hợp với những đánh giá của một cộng sự với trang web Cơ sở Dữ liệu Phim Internet (IMDb). Tổng quát hơn, nghiên cứu đã chứng minh rằng việc đánh giá chỉ 6 bộ phim không có tiếng tăm (trong tốp 500) có thể giúp xác định một khách hàng của Netflix tới 84 phần trăm. Và nếu biết được ngày mà người đó đánh giá phim thì cô ta hoặc anh ta có thể bị chỉ đích danh trong số gần nửa triệu khách hàng thuộc bộ dữ liệu, với độ chính xác 99 phần trăm.

Trong trường hợp AOL, danh tính của người sử dụng được bộc lộ trong nội dung các lệnh tìm kiếm của họ. Trong trường hợp Netflix, danh tính đã được tiết lộ bởi một so sánh các dữ liệu với các nguồn khác. Trong cả hai trường hợp, các công ty đã thất bại và không hề biết dữ liệu lớn đã hỗ trợ phi-vô-danh-hóa tốt như

thế nào. Có hai lý do: chúng ta thu thập nhiều dữ liệu hơn và chúng ta kết hợp nhiều dữ liệu hơn. Paul Ohm, một giáo sư luật tại Đại học Colorado ở Boulder và một chuyên gia về các tổn hại do phi-vô-danh-hóa, giải thích rằng không hề có cách sửa chữa dễ dàng nào cả. Với đủ dữ liệu, không thể ẩn danh tuyệt đối dù cố gắng tới mức nào đi nữa. Tệ hơn, các nhà nghiên cứu gần đây đã chỉ ra rằng không chỉ dữ liệu thông thường mà cả đồ thị xã hội - những kết nối của mọi người với nhau - cũng dễ bị tổn thương vì phi-vô-danh-hóa.

Trong thời đại của dữ liệu lớn, ba chiến lược cốt lõi từ lâu được sử dụng để đảm bảo tính riêng tư - thông báo và xin phép cá nhân, loại ra, và vô danh hóa - đã mất đi phần lớn hiệu quả của chúng. Hiện nay nhiều người sử dụng đã cảm thấy sự riêng tư của họ bị xâm phạm rồi, huống hồ đến lúc việc áp dụng dữ-liệu-lớn trở nên phổ biến hơn.

So với Đông Đức một phần tư thế kỷ trước, việc giám sát đã dễ dàng hơn, rẻ hơn, và mạnh mẽ hơn. Khả năng thu thập dữ liệu cá nhân thường được cấy sâu vào trong các công cụ chúng ta dùng hàng ngày, từ các trang web đến các ứng dụng điện thoại thông minh. Các bộ ghi dữ liệu bên trong hầu hết xe hơi để thu nhận tất cả hoạt động của một chiếc xe vài giây trước lúc túi khí kích hoạt đã được xem như kẻ “làm chứng” chống lại chủ sở hữu xe tại tòa trong các tranh chấp về các sự kiện của tai nạn.

Tất nhiên, khi các doanh nghiệp thu thập dữ liệu để cải thiện hoạt động của họ, chúng ta không cần lo sợ sự giám sát của họ sẽ gây hậu quả như khi bị Stasi nghe trộm. Chúng ta sẽ không bị đi tù nếu Amazon phát hiện chúng ta thích đọc “Little Red Book”. Google sẽ không lưu đày chúng ta chỉ vì chúng ta tìm kiếm từ “Bing”. Các công ty có thể mạnh, nhưng họ không có quyền hạn của nhà nước để ép buộc.

Vì vậy, dù họ không lôi chúng ta khỏi nhà vào giữa đêm, đủ loại

công ty vẫn tích lũy hàng núi thông tin cá nhân liên quan tới tất cả các khía cạnh cuộc sống của chúng ta, chia sẻ nó với những người khác mà chúng ta không hề biết, và sử dụng nó theo những cách mà chúng ta khó có thể tưởng tượng nổi.

Khu vực tư nhân không một mình phô diễn sức mạnh của nó với dữ liệu lớn. Chính phủ cũng đang làm điều đó. Ví dụ Cơ quan An ninh Quốc gia Mỹ (NSA) được cho là chặn và lưu trữ 1,7 tỷ email, cuộc gọi điện thoại, và những liên lạc khác mỗi ngày, theo một điều tra của *Washington Post* trong năm 2010. William Binney, một cựu viên chức NSA, ước tính rằng chính phủ đã thu thập "20.000 tỷ giao dịch" giữa các công dân Mỹ và những người khác - ai gọi ai, gửi email cho ai, chuyển tiền cho ai, vân vân.

Để mang lại ý nghĩa cho tất cả các dữ liệu, Mỹ đang xây dựng những trung tâm dữ liệu khổng lồ, như một cơ sở 1,2 tỷ USD của NSA ở Fort Williams, Utah. Và tất cả cơ quan của chính phủ đang yêu cầu nhiều thông tin hơn so với trước đây, không chỉ riêng các cơ quan bí mật liên quan đến chống khủng bố. Khi việc thu thập mở rộng tới những thông tin như giao dịch tài chính, hồ sơ sức khỏe, và cập nhật trạng thái Facebook, số lượng thông tin được lượm lặt sẽ lớn không thể tưởng tượng nổi. Chính phủ không thể xử lý nhiều dữ liệu như thế. Vậy tại sao lại thu thập nó?

Câu trả lời là cách thức giám sát đã thay đổi trong thời đại dữ liệu lớn. Trong quá khứ, những người điều tra gắn máy vào đường dây điện thoại để tìm hiểu nhiều nhất có thể về một nghi can. Điều quan trọng là đi sâu và tìm hiểu về cá nhân này. Cách tiếp cận hiện đại thì khác. Theo tinh thần của Google hay Facebook, con người là tổng hợp các mối quan hệ xã hội của họ, các tương tác trực tuyến và các kết nối với nội dung. Để điều tra đầy đủ một cá nhân, các nhà phân tích phải nhìn vào khoảng tranh tối tranh sáng rộng nhất có thể của dữ liệu bao quanh con người này - không chỉ những người anh ta quen, mà cả những người quen

của những người quen, và cứ như vậy. Điều này rất khó thực hiện với kỹ thuật trong quá khứ. Ngày nay nó đã dễ dàng hơn bao giờ hết. Và bởi vì chính phủ không bao giờ biết sẽ muốn điều tra kỹ lưỡng ai, nên họ cứ thu thập, lưu trữ, hoặc đảm bảo việc truy cập thông tin, không nhất thiết để theo dõi tất cả mọi người ở mọi thời điểm, nhưng để khi một người nào đó bị nghi ngờ, các nhà chức trách có thể ngay lập tức điều tra thay vì phải bắt đầu thu thập các thông tin từ đầu.

Hoa Kỳ không phải là chính phủ duy nhất tích lũy hàng núi dữ liệu về công dân, cũng không phải là nơi nghiêm túc nhất trong việc này. Tuy nhiên, một vấn đề mới đã xuất hiện với dữ liệu lớn, cũng đáng lo ngại như khả năng các doanh nghiệp và chính phủ biết được thông tin cá nhân của chúng ta: việc sử dụng những dự đoán để đánh giá chúng ta.

Xác suất và hình phạt

John Anderton là chỉ huy một đơn vị cảnh sát đặc nhiệm ở Washington, DC. Một buổi sáng nọ, ông xông vào một ngôi nhà ngoại ô trong khoảnh khắc trước khi Howard Marks, ở trong trạng thái giận dữ điên cuồng, sắp đâm chiếc kéo vào vợ, người mà anh ta thấy trên giường với một gã đàn ông khác. Với Anderton, đó chỉ là một ngày nữa trong trận chiến ngăn chặn tội phạm. “Theo thẩm quyền của Đơn vị Tiền tội phạm của Quận Columbia”, ông đọc thuộc lòng, “tôi bắt giữ anh vì tội giết Sarah Marks trong tương lai, xảy ra vào ngày hôm nay...”. Những cảnh sát khác bắt đầu khống chế Marks trong khi anh ta đang gào lên: “Tôi có *làm* cái gì đâu!”.

Cảnh mở màn của bộ phim *Minority Report* mô tả một xã hội trong đó những dự đoán có vẻ chính xác tới độ cảnh sát bắt giữ người ta vì những tội trạng từ trước khi chúng được thực hiện. Người ta bị giam giữ không phải vì những gì họ đã làm, mà vì những gì họ đang định làm, mặc dù họ không bao giờ thực sự

phạm tội. Bộ phim gán sự thấy trước và việc thực thi pháp luật chặn trước này cho tầm nhìn của ba thần nhãn, chứ không phải cho phân tích dữ liệu. Tuy nhiên, tương lai đáng lo ngại mà *Minority Report* miêu tả là việc phân tích dữ-liệu-lớn không được kiểm soát sẽ mang lại nguy cơ, trong đó những bản án kết tội dựa trên các dự đoán cá nhân hóa của hành vi tương lai.

Chúng ta đã được thấy những mầm mống của điều này. Bảng tạm tha trong hơn một nửa số các tiểu bang của Mỹ sử dụng những dự đoán dựa trên phân tích dữ liệu như một yếu tố để quyết định liệu có nên thả một ai đó khỏi nhà tù hay giam giữ anh ta. Ngày càng có nhiều nơi ở Hoa Kỳ - từ các phân khu ở Los Angeles đến các thành phố như Richmond, Virginia - áp dụng "chính sách tiên đoán": dùng phân tích dữ-liệu-lớn để chọn những đường phố, nhóm và cá nhân phải bị giám sát thêm, đơn giản vì một thuật toán chỉ ra là họ có nhiều khả năng phạm tội.

Tại thành phố Memphis, Tennessee, một chương trình gọi là Blue CRUSH (Giảm Tội phạm bằng cách Sử dụng Lịch sử Thống kê) cung cấp cho cảnh sát tương đối chính xác các khu vực cần quan tâm về địa điểm (một vài khối phố) và thời gian (một vài giờ trong một ngày đặc biệt của tuần). Hệ thống dường như giúp lực lượng thực thi pháp luật phân bổ nguồn lực khan hiếm của họ tốt hơn. Từ khi chương trình được triển khai vào năm 2006, những vụ phạm tội với tài sản lớn và các hành vi bạo lực đã giảm một phần tư (mặc dù tất nhiên, điều này không nói lên được gì về quan hệ nhân quả, cũng không có gì để cho biết rằng sự sụt giảm là nhờ Blue CRUSH).

Ở Richmond, bang Virginia, cảnh sát lập tương quan dữ liệu tội phạm với các bộ dữ liệu khác, ví dụ thông tin khi nào các công ty lớn trong thành phố trả lương cho nhân viên của họ, những ngày diễn ra các buổi hòa nhạc hoặc các sự kiện thể thao. Làm như vậy đã xác nhận và đôi khi tinh lọc những nghi ngờ của cảnh sát về xu hướng tội phạm. Ví dụ cảnh sát Richmond một thời gian dài

cảm nhận có một sự tăng về tội phạm bạo lực tiếp sau các triển lãm súng. Phân tích dữ liệu lớn đã chứng tỏ họ đúng, nhưng không hoàn toàn: sự tăng thường xảy ra hai tuần sau đó, chứ không phải ngay lập tức sau những sự kiện này.

Các hệ thống như trên hướng đến việc phòng ngừa tội phạm bằng cách dự đoán, mục tiêu cuối cùng là đến tận cấp độ cá nhân - những kẻ có thể gây ra chúng. Điều này cho thấy khả năng sử dụng dữ liệu lớn cho một mục đích mới: để ngăn chặn tội phạm khỏi xảy ra.

Một dự án nghiên cứu trực thuộc Bộ An ninh nội địa Hoa Kỳ (DHS) được gọi là FAST (Công nghệ Sàng lọc Thuộc tính Tương lai) cố xác định những kẻ có nguy cơ trở thành khủng bố bằng cách theo dõi các dấu hiệu sống của cá nhân, ngôn ngữ cơ thể, và các mô hình sinh lý khác. Ý tưởng ở đây là việc khảo sát hành vi của con người có thể phát hiện được ý định gây hại của họ. Trong các thử nghiệm, hệ thống chính xác đến 70 phần trăm, theo DHS. (Điều này có nghĩa là gì thì không rõ. Phải chăng các đối tượng tham gia nghiên cứu giả vờ làm khủng bố để xem "ý định xấu" của họ có được phát hiện?) Mặc dù các hệ thống này dường như còn phôi thai, nhưng vấn đề là lực lượng thực thi pháp luật xem chúng rất nghiêm túc.

Ngăn chặn một tội phạm để nó không xảy ra dường như là một viễn cảnh hấp dẫn. Chẳng phải việc ngăn chặn các vi phạm trước khi chúng xảy ra là tốt hơn nhiều so với xử phạt các thủ phạm sau đó hay sao? Chẳng phải việc chặn các tội ác đem lại lợi ích không chỉ cho những người có thể là nạn nhân của chúng, mà còn cho toàn thể xã hội hay sao?

Nhưng đó là một con đường nguy hiểm. Nếu thông qua dữ liệu lớn để dự đoán được ai có thể phạm tội trong tương lai, chúng ta hẳn sẽ không bằng lòng với việc chỉ đơn giản ngăn chặn tội phạm xảy ra, mà còn muốn trừng phạt kẻ có thể là thủ phạm

nữa. Điều đó hợp logic. Nếu chúng ta chỉ bước vào can thiệp để ngăn chặn hành động bất hợp pháp khỏi diễn ra, kẻ được xem là thủ phạm có thể sẽ thử lại mà không bị trừng phạt. Ngược lại, bằng cách sử dụng phân tích dữ liệu lớn để bắt hắn phải chịu trách nhiệm đối với những hành vi (tương lai) của mình, chúng ta có thể ngăn được hắn và cả những kẻ khác nữa.

Sự trừng phạt dựa trên dự đoán như vậy có vẻ là một bước cải thiện so với những biện pháp mà chúng ta đã chấp nhận. Việc ngăn chặn hành vi không lành mạnh, nguy hiểm, hoặc rủi ro là một nền tảng của xã hội hiện đại. Chúng ta đã gây khó khăn cho việc hút thuốc để ngăn ngừa bệnh ung thư phổi, chúng ta yêu cầu thắt dây an toàn để ngăn ngừa tử vong trong tai nạn xe hơi, chúng ta không cho phép hành khách lên máy bay với súng để tránh cướp. Những biện pháp phòng ngừa như vậy hạn chế sự tự do của chúng ta, nhưng nhiều người xem chúng như cái giá nhỏ phải trả để tránh được tác hại nghiêm trọng hơn nhiều.

Trong nhiều trường hợp, phân tích dữ liệu đã được sử dụng nhân danh việc phòng ngừa. Nó được sử dụng để gộp chúng ta vào nhóm của những người giống chúng ta, và chúng ta thường được đặc trưng hóa theo đó. Bảng tính toán bảo hiểm lưu ý rằng những người đàn ông hơn 50 tuổi dễ bị ung thư tuyến tiền liệt, vì vậy các thành viên của nhóm này có thể phải trả nhiều hơn cho bảo hiểm y tế ngay cả khi họ không bao giờ mắc bệnh ung thư tuyến tiền liệt. Nhóm học sinh trung học với điểm cao ít có khả năng bị tai nạn xe hơi - vì vậy một số bạn học kém hơn của họ phải đóng bảo hiểm cao hơn. Những cá nhân với một số đặc điểm nào đó là đối tượng kiểm tra chặt chẽ hơn khi họ đi qua an ninh sân bay.

Đó là ý tưởng đằng sau việc “lập hồ sơ” trong thế giới dữ-liệu-nhỏ ngày nay. Tìm một liên hợp chung trong dữ liệu, xác định một nhóm người để áp dụng vào, và sau đó đặt những người này dưới sự giám sát bổ sung. Đó là một quy tắc khái quát áp dụng cho tất

cả mọi người trong nhóm. Tất nhiên phương pháp này có nhược điểm nghiêm trọng. Nếu được sử dụng không đúng, nó có thể dẫn tới không chỉ sự phân biệt đối xử với những nhóm nhất định mà còn cả “phạm tội vì đồng lõa”.

Ngược lại, dự báo dữ liệu lớn về con người lại khác. Trong khi các dự báo ngày nay về hành vi có thể xảy ra - được tìm thấy trong những thứ như phí bảo hiểm hoặc điểm số tín dụng - thường căn cứ vào rất nhiều yếu tố được dựa trên một mô hình của vấn đề đang xét (chẳng hạn vấn đề về sức khỏe trước đây hay lịch sử trả tiền vay nợ), với phân tích phi nhân quả của dữ liệu lớn, chúng ta thường chỉ đơn giản xác định các yếu tố dự báo phù hợp nhất từ biển thông tin.

Quan trọng nhất, sử dụng dữ liệu lớn, chúng ta hy vọng sẽ xác định được các cá nhân cụ thể chứ không phải là các nhóm, điều này giải thoát chúng ta khỏi thiếu sót của lập hồ sơ làm cho mỗi nghi ngờ được dự đoán trở thành một trường hợp của tội đồng lõa. Trong một thế giới dữ-liệu-lớn, ai đó với một cái tên Ả Rập, trả tiền mặt cho một chiếc vé một chiều hạng nhất, có thể không còn phải bị kiểm tra bổ sung tại sân bay nếu các dữ liệu khác chứng tỏ chắc chắn rằng anh ta không phải là một tên khủng bố. Với dữ liệu lớn chúng ta có thể thoát khỏi sự bó buộc vào đặc điểm của cả nhóm, và thay vào đó có thể đưa ra nhiều dự đoán chi tiết cho cá nhân hơn.

Triển vọng của dữ liệu lớn là chúng ta có thể làm những gì mình đã làm trong suốt thời gian qua - lập hồ sơ - nhưng khiến nó tốt hơn, ít phân biệt đối xử hơn, và cá nhân hóa nhiều hơn. Nghe có vẻ chấp nhận được nếu mục đích chỉ đơn giản là để ngăn chặn những hành động không mong muốn. Nhưng nó trở nên rất nguy hiểm nếu chúng ta sử dụng các dự đoán dữ-liệu-lớn để quyết định xem liệu ai đó là có tội và phải bị trừng phạt vì hành vi chưa xảy ra.

Ý tưởng về xử phạt chỉ dựa trên các khuynh hướng là một ý tưởng tồi tệ. Để buộc tội một người vì các hành vi có thể xảy ra trong tương lai là phủ nhận nền tảng rất cơ bản của công lý: người này phải làm điều gì đó trước khi chúng ta có thể buộc anh ta chịu trách nhiệm về nó. Xét cho cùng, nghĩ đến những điều xấu không phải là bất hợp pháp, nhưng thực hiện chúng lại là bất hợp pháp. Đó là một nguyên lý cơ bản của xã hội chúng ta rằng trách nhiệm cá nhân gắn liền với sự lựa chọn cá nhân của hành động. Nếu một người bị buộc phải dùng súng để bảo vệ sự an toàn của mình, anh ta không có sự lựa chọn nào khác và do đó không bị buộc chịu trách nhiệm.

Nếu các dự đoán dữ-liệu-lớn là hoàn hảo, nếu các thuật toán có thể đoán trước tương lai của chúng ta với độ rõ nét hoàn hảo, chúng ta sẽ không còn quyền lựa chọn để hành động trong tương lai. Chúng ta sẽ hành xử đúng như được dự đoán. Nếu các dự đoán hoàn hảo là khả thi, chúng sẽ gạt bỏ ý chí của con người - khả năng của chúng ta để tự do sống cuộc đời mình. Nhưng trớ trêu thay, bằng cách tước đoạt khỏi chúng ta sự lựa chọn, chúng cũng miễn xá cho chúng ta khỏi bất kỳ trách nhiệm nào.

Tất nhiên dự đoán hoàn hảo là không thể. Thay vào đó, phân tích dữ-liệu-lớn sẽ dự đoán rằng đối với một cá nhân cụ thể, một hành vi cụ thể trong tương lai có một xác suất nhất định. Hãy xét nghiên cứu được tiến hành bởi Richard Berk, một giáo sư về thống kê và tội phạm học tại Đại học Pennsylvania. Ông khẳng định phương pháp của mình có thể dự đoán liệu một người được cho tại ngoại sẽ tham gia vào một vụ giết người (giết hoặc bị giết). Trong thông tin đầu vào, ông sử dụng nhiều tham biến gắn với trường hợp cụ thể, trong đó có lý do bị tù và ngày vi phạm lần đầu, nhưng cũng sử dụng dữ liệu nhân khẩu học như tuổi tác và giới tính. Berk cho thấy ông có thể dự báo một vụ giết người trong tương lai trong số những người được tạm tha với xác suất tối thiểu là 75 phần trăm. Con số đó không hề thấp. Tuy nhiên, nó cũng có nghĩa là nếu các hội đồng xét xử dựa vào phân tích

của Berk thì họ sẽ sai lầm tới một phần tư số trường hợp.

Nhưng vấn đề cốt lõi khi dựa vào những dự đoán như vậy không phải là nó đưa xã hội tới rủi ro. Rắc rối cơ bản là với một hệ thống như vậy, chúng ta chủ yếu trừng phạt mọi người trước khi họ làm điều xấu. Và bằng cách can thiệp trước khi họ hành động (ví dụ bằng cách từ chối tạm tha nếu các dự đoán cho thấy có một xác suất cao là họ sẽ giết người), chúng ta không bao giờ biết liệu họ có phạm tội được dự đoán. Dù không chấp nhận vận số, nhưng chúng ta lại buộc các cá nhân phải chịu trách nhiệm về những gì mà dự đoán của chúng ta tiết lộ rằng họ sẽ thực hiện. Những dự đoán như vậy không bao giờ có thể bác bỏ được. Điều này phủ nhận ý tưởng tối thượng về giả định vô tội, nguyên tắc mà hệ thống pháp luật của chúng ta, cũng như ý thức của chúng ta về sự công bằng, vẫn dựa vào. Và nếu buộc mọi người chịu trách nhiệm về những hành vi tương lai được dự đoán, mà họ có thể không bao giờ phạm phải, thì chúng ta cũng phủ nhận rằng con người có một năng lực cho sự lựa chọn mang tính đạo đức.

Điểm quan trọng ở đây không chỉ liên quan đến an ninh trật tự. Mối hiểm họa trải rộng hơn nhiều, bao gồm tất cả các lĩnh vực của xã hội, tất cả các trường hợp phán quyết của con người trong đó những dự đoán dữ-liệu-lớn được sử dụng để quyết định xem một người có mắc tội với những hành vi tương lai hay không. Chúng bao gồm tất cả mọi thứ, từ quyết định của một công ty thải hồi một nhân viên, một bác sĩ từ chối phẫu thuật một bệnh nhân, đến một người vợ/chồng nộp đơn ly dị.

Có lẽ với một hệ thống như vậy, xã hội sẽ được an toàn hơn và hiệu quả hơn, nhưng một phần thiết yếu của những gì khiến chúng ta là con người - khả năng lựa chọn các hành động của mình và phải chịu trách nhiệm về chúng - sẽ bị phá hủy. Dữ liệu lớn sẽ trở thành một công cụ để tập thể hóa lựa chọn của con người và từ bỏ ý chí tự do trong xã hội của chúng ta.

Tất nhiên, dữ liệu lớn cung cấp rất nhiều lợi ích. Điều biến nó thành một thứ vũ khí của phi nhân hóa chỉ là một khiếm khuyết, không phải của chính bản thân dữ liệu lớn, mà của những cách thức chúng ta sử dụng các dự đoán của nó. Điểm bất cập chính là buộc con người phải chịu tội, xuất phát từ những dự đoán dữ-liệu-lớn dựa trên mối tương quan nhưng lại đưa ra những quyết định có quan hệ nhân quả về trách nhiệm cá nhân.

Dữ liệu lớn rất hữu ích để hiểu được nguy cơ hiện tại và tương lai, và để điều chỉnh hành động của chúng ta một cách phù hợp. Nhưng dữ liệu lớn không cho chúng ta bất cứ điều gì về quan hệ nhân quả. Việc gán "tội lỗi" - tội lỗi cá nhân - đòi hỏi rằng những người mà chúng ta phán quyết đã chọn một hành động cụ thể. Quyết định của họ phải là nguyên nhân cho hành động. Chính vì dữ liệu lớn được dựa trên các mối tương quan, nên nó là công cụ hoàn toàn không phù hợp để giúp chúng ta phán quyết quan hệ nhân quả và do đó khép tội cho cá nhân.

Vấn đề là con người chủ yếu nhìn thế giới qua lăng kính của nhân quả. Do đó dữ liệu lớn luôn có nguy cơ bị lạm dụng cho các mục đích quan hệ nhân quả, bị gắn liền với những lăng kính màu hồng, cho rằng sự phán xét của chúng ta có thể hiệu quả nhiều hơn đến thế nào, chỉ cần ta được trang bị những dự đoán dữ-liệu-lớn.

Nó đúng là con dốc trơn thuần túy - trượt thẳng đến xã hội được miêu tả trong bộ phim *Minority Report*, một thế giới mà trong đó sự lựa chọn cá nhân và ý chí tự do bị loại bỏ, định hướng đạo đức cá nhân của chúng ta bị thay thế bởi các thuật toán dự đoán, và các cá nhân phải đối mặt với mũi giùi không hề bị ngăn trở của những sắc lệnh tập thể. Nếu được sử dụng như vậy thì dữ liệu lớn đe dọa sẽ giam cầm chúng ta - có lẽ theo nghĩa đen - trong nhà tù xác suất.

Độc tài dữ liệu

Dữ liệu lớn làm xói mòn sự riêng tư và đe dọa tự do. Nhưng dữ liệu lớn cũng làm trầm trọng thêm một vấn đề rất cũ: sự tin cậy vào những con số dù chúng dễ sai hơn chúng ta tưởng nhiều. Không gì nhấn mạnh những hậu quả méo mó của phân tích dữ liệu hơn câu chuyện của Robert McNamara.

McNamara là người của những con số. Được bổ nhiệm làm Bộ trưởng Quốc phòng Mỹ vào đầu những năm 1960, ông yêu cầu nhận được dữ liệu về tất cả mọi thứ có thể. Chỉ bằng cách áp dụng sự chặt chẽ của thống kê, ông tin rằng những người ra quyết định có thể hiểu được một tình huống phức tạp và đưa ra những lựa chọn đúng đắn. Thế giới trong quan niệm của ông là một khối thông tin hỗn độn, nếu được miêu tả, biểu lộ, phân định, và định lượng thì có thể được chế ngự bởi bàn tay con người và sẽ phục vụ ý muốn của con người. McNamara muốn tìm kiếm Sự Thật từ dữ liệu. Và trong những con số được gửi về cho ông là "số xác chết".

McNamara biểu lộ đam mê các con số khi còn là một sinh viên tại Trường Kinh doanh của Đại học Harvard và sau đó là giáo sư dự khuyết trẻ nhất của trường ở tuổi 24. Ông đã áp dụng khoa học chặt chẽ này trong Chiến tranh Thế giới Thứ hai khi là thành viên của nhóm ưu tú Lầu Năm Góc tên là Ban Điều khiển Thống kê, đưa quy trình ra-quyết-định-dựa-trên-dữ-liệu vào một trong những bộ máy quan liêu lớn nhất thế giới. Trước đó, quân đội vẫn mù mờ thông tin. Ví dụ họ không biết loại, số lượng, hay vị trí của các phụ tùng máy bay. Dữ liệu đã đến để giải cứu. Chỉ mỗi việc mua sắm vũ khí hiệu quả hơn đã cắt giảm được $3,6 tỷ trong năm 1943. Chiến tranh hiện đại là về sự phân bổ hiệu quả các nguồn lực, và do vậy công việc của nhóm nghiên cứu là một thành công tuyệt vời.

Khi chiến tranh kết thúc, nhóm quyết định gắn bó với nhau và đóng góp những kỹ năng của họ cho các công ty Mỹ. Công ty Ford Motor đang gặp khó khăn, và một Henry Ford II tuyệt vọng

đã trao dây cương cho họ. Giống như lúc tham gia nhóm nghiên cứu cho quân đội, họ cũng chẳng hề biết gì về chế tạo xe hơi. Tuy nhiên, những người được mệnh danh là những “Đứa trẻ Thần đồng” này đã xoay chuyển công ty.

McNamara đã thăng tiến nhanh chóng lên các cấp bậc, luôn nhanh chóng đưa ra một điểm dữ liệu cho mỗi tình huống. Các quản lý xí nghiệp bực bội cung cấp những con số mà ông yêu cầu - cho dù chúng chính xác hay không. Khi một chỉ thị đưa xuống rằng tất cả hàng tồn kho của một mô hình xe hơi phải được sử dụng trước khi một mô hình mới có thể bắt đầu sản xuất, các quản lý phân xưởng tức tối đổ các phụ tùng thừa xuống một con sông gần đó. Lãnh đạo tại trụ sở trung ương đồng ý phê duyệt khi các thợ cả báo cáo những con số xác nhận rằng chỉ thị đã được tuân thủ. Nhưng ở nhà máy người ta vẫn nói đùa rằng họ có thể đi trên mặt nước - trên những mảnh sắt gỉ của những chiếc xe đời 1950 và 1951.

McNamara là hình ảnh thu nhỏ của người quản lý giữa thế kỷ XX, người điều hành siêu hợp lý tin tưởng vào các con số thay vì cảm tính, và người có thể áp dụng những kỹ năng định lượng của mình vào bất cứ ngành công nghiệp nào mà ông ta tiếp cận. Năm 1960, ông trở thành chủ tịch của Ford, một vị trí mà ông chỉ nắm giữ vài tuần trước khi Tổng thống Kennedy bổ nhiệm ông làm Bộ trưởng Quốc phòng.

Khi cuộc chiến Việt Nam leo thang và Hoa Kỳ gửi nhiều quân đội hơn, cách thức để đo lường sự tiến triển là bằng số lượng kẻ thù bị giết chết. Số liệu được công bố hàng ngày trên báo chí. Với những kẻ ủng hộ chiến tranh, nó là bằng chứng của sự tiến triển, còn với những ai phản đối, nó là bằng chứng của sự vô đạo đức. Số xác chết là điểm dữ liệu đã định nghĩa một thời đại.

Năm 1977, hai năm sau khi chiếc trực thăng cuối cùng cất cánh khỏi nóc tòa đại sứ quán Mỹ ở Sài Gòn, một tướng quân đội về

hưu, Douglas Kinnard, công bố một khảo sát mang tính bước ngoặt về quan điểm của các vị tướng. Được lấy tên là *The War Managers*, cuốn sách tiết lộ vũng lầy của định lượng. Chỉ hai phần trăm các tướng lĩnh của Mỹ xem số liệu xác chết là một cách hợp thức để đo lường sự tiến triển. Khoảng hai phần ba nói nó thường bị thổi phồng. “Một sự giả mạo - hoàn toàn vô giá trị”, một tướng nhận định. “Thường là những lời nói dối trắng trợn”, một người khác viết. “Chúng bị phóng đại hết cỡ bởi nhiều đơn vị, chủ yếu vì sự quan tâm quá đáng của những kẻ như McNamara”, một người thứ ba thẳng thừng nói.

Như những công nhân tại nhà máy Ford đã vứt các phụ tùng động cơ xuống sông, các sĩ quan trẻ đôi khi báo cáo cho cấp trên của họ những con số ấn tượng để giữ vai trò chỉ huy hoặc thúc đẩy sự nghiệp riêng. McNamara và những tướng tá xung quanh ông ta đã tin tưởng vào các con số, mê đắm chúng.

Việc sử dụng, lạm dụng và gian lận trong dữ liệu của quân đội Mỹ trong chiến tranh ở Việt Nam là một bài học đáng lo ngại về những hạn chế của thông tin trong thời đại dữ liệu nhỏ, một bài học phải được lưu ý khi thế giới tiến tới thời đại dữ-liệu-lớn. Chất lượng của các dữ liệu nền tảng có thể nghèo nàn. Nó có thể bị sai lệch. Nó có thể bị phân tích sai hoặc sử dụng một cách sai lạc. Và thậm chí tệ hại hơn, dữ liệu có thể không phản ánh được những gì nó nhắm tới để định lượng.

Chúng ta hiểu về thuật ngữ “độc tài dữ liệu” hơn mình tưởng. Thuật ngữ này nghĩa là để cho dữ liệu chi phối mình theo những cách thức có thể gây nhiều điều thiệt hại hơn là điều tốt. Mối đe dọa nằm ở chỗ chúng ta để cho chính mình bị ràng buộc một cách vô thức bởi kết quả của các phân tích ngay cả khi có những căn cứ hợp lý để nghi ngờ điều gì đó là không ổn. Hoặc chúng ta sẽ bị ám ảnh bởi việc thu thập dữ kiện và số liệu, thu thập chỉ để thu thập. Hoặc chúng ta sẽ gán cho dữ liệu một mức độ chân thật mà nó không xứng đáng được nhận.

Khi nhiều khía cạnh hơn của cuộc sống được dữ liệu hóa, giải pháp mà các nhà hoạch định chính sách và doanh nhân bắt đầu muốn nhắm tới là có được nhiều dữ liệu hơn. “Chúng ta tin ở Chúa - còn tất cả những thứ khác thì mang đến dữ liệu”, đây là câu thần chú của các nhà quản lý hiện đại, được nghe vang vọng khắp Thung Lũng Silicon, trên các sàn nhà máy, và dọc hành lang của các cơ quan chính phủ. Ngụ ý thì lành mạnh, nhưng người ta có thể dễ dàng bị đánh lừa bởi dữ liệu.

Giáo dục dường như trượt dốc? Hãy thúc đẩy các bài kiểm tra chuẩn hóa để đo lường hiệu suất và trừng phạt các giáo viên hoặc trường học nào không đạt. Cho dù các bài kiểm tra có thực sự nắm bắt được khả năng của học sinh hay không, chất lượng giảng dạy, hoặc nhu cầu của một lực lượng lao động sáng tạo, hiện đại và có khả năng thích ứng vẫn là một câu hỏi mở - nhưng là một điều mà dữ liệu không thừa nhận.

Muốn ngăn chặn khủng bố? Hãy tạo những lớp danh sách giám sát và cấm bay để kiểm soát bầu trời. Nhưng liệu những bộ dữ liệu như vậy có cung cấp nổi sự bảo vệ mà chúng hứa hẹn hay không thì còn phải bàn lại. Trong một sự cố nổi tiếng, cố Thượng nghị sĩ Ted Kennedy của bang Massachusetts, Mỹ đã bị “sa lưới” bởi danh sách cấm bay, bị chặn lại và thẩm vấn, chỉ đơn giản vì có tên giống một người trong cơ sở dữ liệu.

Những người làm việc với dữ liệu có một cách diễn đạt cho những vấn đề như vậy: “rác vào, rác ra”. Trong một số trường hợp nhất định, nguyên nhân nằm ở chất lượng của các thông tin cơ bản. Tuy nhiên nó thường do sự lạm dụng kết quả phân tích. Với dữ liệu lớn, những vấn đề này có thể xuất hiện thường xuyên hơn hoặc có những hậu quả lớn hơn.

Google, như chúng ta đã chỉ rõ trong nhiều ví dụ, thực hiện tất cả mọi thứ theo dữ liệu. Chiến lược này rõ ràng đã dẫn đến nhiều thành công. Nhưng thỉnh thoảng nó cũng làm cho công ty lao

đao. Các đồng sáng lập của công ty, Larry Page và Sergey Brin, từ lâu đã kiên quyết yêu cầu được biết điểm thi SAT và điểm trung bình khi tốt nghiệp đại học của các ứng cử viên. Trong suy nghĩ của họ, con số đầu thể hiện tiềm năng và con số thứ hai thể hiện thành tích. Cả các nhà quản lý tài năng ở độ tuổi bốn mươi được tuyển dụng cũng bị hối thúc cung cấp các điểm số, và họ hoàn toàn bối rối về điều đó. Công ty thậm chí vẫn tiếp tục yêu cầu cung cấp điểm số, rất lâu sau khi các nghiên cứu nội bộ của nó cho thấy không có mối tương quan nào giữa điểm số và hiệu suất công việc.

Google đáng ra phải biết nhiều hơn, để không bị sự quyến rũ sai lệch của dữ liệu lôi cuốn. Cách đo lường này mang lại rất ít cơ hội cho thay đổi trong cuộc sống của một con người. Nó thất bại trong việc tính đến kiến thức chứ không phải sự thông minh sách vở. Và nó có thể không phản ánh năng lực của những người từ các ngành nhân văn, nơi hiểu biết có thể khó được định lượng hơn trong ngành khoa học và kỹ thuật. Nỗi ám ảnh đó của Google với dữ liệu liên quan đến nhân sự là đặc biệt lạ lùng, nếu xét rằng những người sáng lập công ty là những sản phẩm của trường phái Montessori, trong đó nhấn mạnh việc học tập, chứ không phải điểm số. Và nó lặp lại những sai lầm của những cường quốc công nghệ quá khứ, coi trọng hồ sơ hơn các khả năng thực tế của ứng viên. Liệu Larry và Sergey, từng bỏ ngang khi làm nghiên cứu sinh tiến sĩ, có giành được một cơ hội để trở thành các nhà quản lý tại Bell Labs huyền thoại? Theo các tiêu chuẩn của Google, Bill Gates, cũng như Mark Zuckerberg, và Steve Jobs đều sẽ không được thuê, vì không có bằng đại học.

Sự phụ thuộc của công ty vào dữ liệu đôi khi có vẻ bị thổi phồng. Marissa Mayer, khi là một trong những giám đốc điều hành hàng đầu của Google, một lần đã lệnh cho nhân viên thử 41 sắc màu xanh để xem sắc nào được ưa chuộng sử dụng nhiều hơn, nhằm xác định màu sắc của một thanh công cụ trên trang web. Sự sùng bái dữ liệu của Google đã tới mức cực đoan. Nó thậm chí

còn gây ra cuộc nổi loạn.

Năm 2009, nhà thiết kế hàng đầu của Google, Douglas Bowman, đã bỏ đi trong một cơn tức giận vì ông không thể chịu được áp lực phải lượng hóa liên tục tất cả mọi thứ. “Tôi đã có một cuộc tranh luận gần đây về việc liệu một đường biên nên rộng 3, 4 hay 5 điểm ảnh, và được yêu cầu phải chứng minh đề nghị của mình. Tôi không thể làm việc trong một môi trường như vậy”, ông đã viết trên blog để thông báo việc từ chức của mình. “Khi một công ty gồm toàn các kỹ sư, nó sẽ bám lấy kỹ thuật để giải quyết các vấn đề. Mỗi quyết định đều phải quy về một bài toán logic đơn giản. Cuối cùng dữ liệu trở thành một thứ để chống đỡ cho mọi quyết định, làm tê liệt công ty”.

Sự sáng suốt không phụ thuộc vào dữ liệu. Steve Jobs có thể liên tục cải thiện máy tính xách tay Mac trong nhiều năm trên cơ sở các báo cáo thực địa, nhưng ông đã sử dụng trực giác của mình, chứ không phải dữ liệu, để khởi động iPod, iPhone, và iPad. Ông đã dựa vào giác quan thứ sáu. Jobs từng có một phát biểu nổi tiếng, khi trả lời câu hỏi của một phóng viên vì sao Apple không làm nghiên cứu thị trường trước khi phát hành iPad: “Chuyện biết mình muốn gì không phải là việc của người tiêu dùng”.

Trong cuốn sách *Seeing Like a State*, nhà nhân chủng học James Scott của Đại học Yale ghi lại những cách thức mà các chính phủ, khi tôn sùng định lượng hóa và dữ liệu, rốt cuộc chỉ làm cho cuộc sống của người dân thành khốn khổ chứ không trở nên tốt hơn. Họ sử dụng bản đồ nhằm xác định cách tổ chức lại các cộng đồng thay vì tìm hiểu mọi thứ về con người trên mặt đất. Họ sử dụng những bảng dài dữ liệu về thu hoạch để quyết định tập thể hóa nông nghiệp mà không biết một chút gì về nuôi trồng. Việc sử dụng dữ liệu, theo quan điểm của Scott, thường là để trao quyền cho kẻ mạnh.

Đây là độc tài dữ liệu. Và sự ngạo mạn tương tự đã đẩy Hoa Kỳ

leo thang trong chiến tranh Việt Nam, một phần dựa trên cơ sở của số xác chết, chứ không phải những quyết định dựa trên các số liệu có ý nghĩa hơn. “Đúng là không phải mọi tình huống phức tạp về con người đều có thể được quy kết hoàn toàn thành các đường trên một đồ thị, hoặc thành các tỷ lệ phần trăm trên một biểu đồ, hoặc các con số trên một bảng thống kê”, McNamara phát biểu vào năm 1967, khi các cuộc biểu tình trong nước đang tăng. “Nhưng tất cả thực tế đều có thể được lý giải. Và không định lượng những gì có thể được định lượng cũng tức là bằng lòng với việc không xem xét đầy đủ các lý do”. Tuy nhiên vấn đề là sử dụng đúng các dữ liệu đúng, chứ không phải chỉ thu thập cho có.

Robert Strange McNamara chuyển sang phụ trách Ngân hàng Thế giới trong suốt những năm 1970, sau đó tô điểm mình như một con chim bồ câu trong những năm 1980. Sau này ông cho xuất bản cuốn hồi ký *In Retrospect* chỉ trích lối tư duy và những quyết định của chính mình trong vai trò Bộ trưởng Quốc phòng. “Chúng tôi đã sai, quá sai”, ông viết. Nhưng đó là ông đề cập đến chiến lược rộng của chiến tranh. Về vấn đề dữ liệu, và đặc biệt là những con số thương vong, McNamara vẫn không ăn năn. Ông ta thừa nhận nhiều số liệu thống kê đã “lừa dối hoặc sai sót”. “Nhưng những gì có thể đếm, bạn cần phải đếm. Thiệt hại về người là một trong số đó”. McNamara qua đời năm 2009 ở tuổi 93 - một người thông minh nhưng không khôn ngoan.

Dữ liệu lớn có thể lôi kéo chúng ta phạm tội lỗi của McNamara: trở nên quá gắn chặt với dữ liệu, và bị ám ảnh bởi sức mạnh và triển vọng của nó tới mức chúng ta không đánh giá đúng các hạn chế của nó. Thử nhìn lại Xu hướng Dịch cúm của Google. Hãy xem xét một tình huống, không hoàn toàn viển vông, trong đó một chủng nguy hiểm chết người của bệnh cúm dữ dội lan khắp toàn quốc. Các chuyên gia y khoa sẽ biết ơn khả năng dự báo trong thời gian thực các điểm nóng nhất bằng cách rà soát các truy vấn tìm kiếm. Họ sẽ biết được nơi nào phải can thiệp để giúp

đỡ.

Nhưng giả sử rằng trong một thời điểm của cuộc khủng hoảng, các nhà lãnh đạo chính trị cho rằng chỉ biết nơi nào căn bệnh này có thể trở nên tồi tệ hơn và cố gắng loại trừ nó đi là không đủ. Vì vậy, họ yêu cầu một cuộc cách ly trên diện rộng - không phải với tất cả mọi người trong những vùng này, vì không cần thiết và quá rộng. Dữ liệu lớn cho phép chúng ta tập trung hơn. Thế nên việc cách ly chỉ áp dụng với những cá nhân thực hiện các lệnh tìm kiếm nào cho thấy họ nhiều khả năng nhiễm bệnh nhất. Ở đây chúng ta có dữ liệu về họ để lọc ra. Các đặc vụ của liên bang, nắm trong tay danh sách các địa chỉ IP và thông tin GPS của điện thoại di động, sẽ gom những người này vào các trung tâm cách ly.

Dù kịch bản này hợp lý với một số người, nó lại hoàn toàn sai. Các mối tương quan không có nghĩa là quan hệ nhân quả. Những người này có thể bị hoặc có thể không bị cúm. Họ cần phải được kiểm tra. Họ sẽ trở thành các tù nhân của một dự đoán, nhưng quan trọng hơn, họ sẽ là những nạn nhân của một cách nhìn dữ liệu thiển cận, không hiểu đủ ý nghĩa thực sự của thông tin. Vấn đề nằm ở chỗ một số thuật ngữ tìm kiếm nào đó *có tương quan* với sự bùng nổ bệnh dịch - nhưng mối tương quan có thể tồn tại vì những tình huống như các đồng nghiệp khỏe mạnh nghe hắt hơi trong văn phòng và lên mạng để tìm hiểu cách để tự bảo vệ mình, chứ không phải vì chính những người tìm kiếm bị mắc bệnh.

Mặt tối của dữ liệu lớn

Như chúng ta đã thấy, dữ liệu lớn cho phép giám sát cuộc sống của chúng ta nhiều hơn, trong khi nó khiến một số biện pháp pháp lý để bảo vệ sự riêng tư hầu như trở nên lỗi thời. Cũng đáng lo ngại khi các dự đoán dữ-liệu-lớn về cá nhân có thể được sử dụng để trừng phạt công dân vì những khuynh hướng của họ,

chứ không phải vì những hành động của họ. Điều này phủ nhận ý chí tự do và làm xói mòn phẩm giá con người.

Đồng thời, có một nguy cơ thực sự rằng các lợi ích của dữ liệu lớn sẽ lôi kéo người ta áp dụng các kỹ thuật không hoàn toàn thích hợp với họ, hoặc tạo cảm giác quá tin vào các kết quả phân tích. Khi các dự đoán dữ-liệu-lớn được cải thiện, việc sử dụng chúng sẽ càng trở nên hấp dẫn, thúc đẩy một nỗi ám ảnh về dữ liệu vì nó có thể làm được rất nhiều thứ. Đó là lời nguyền của McNamara và là bài học mà câu chuyện về ông ta lưu giữ.

Trong chương tiếp theo, chúng ta sẽ xem xét những cách thức có thể kiểm soát được dữ liệu lớn, thay vì bị nó kiểm soát.

9. KIỂM SOÁT

NHỮNG THAY ĐỔI TRONG CÁCH THỨC chúng ta sản xuất và tương tác với thông tin dẫn đến những thay đổi trong các quy tắc chúng ta sử dụng để quản lý chính mình, và trong các giá trị mà xã hội phải bảo vệ. Hãy xem xét một ví dụ từ cuộc đại hồng thủy dữ liệu trước đây, được giải phóng nhờ công nghệ in ấn.

Trước khi Johannes Gutenberg phát minh ra công nghệ xếp chữ khoảng năm 1450, việc truyền bá ý tưởng ở phương Tây phần lớn bị giới hạn trong các kết nối cá nhân. Sách chủ yếu giới hạn trong các thư viện của tu viện, được trông coi nghiêm ngặt bởi các tu sĩ đại diện cho Giáo hội Công giáo để bảo vệ và bảo tồn sự thống trị của nó. Bên ngoài Giáo Hội, sách cực kỳ hiếm. Một số trường đại học đã thu thập được chỉ vài chục hoặc có thể vài trăm cuốn sách. Đại học Cambridge bắt đầu từ thế kỷ XV với chỉ 122 pho sách.

Trong vòng một vài thập kỷ sau phát minh của Gutenberg, công nghệ in ấn của ông đã được nhân rộng trên khắp châu Âu, khiến việc sản xuất hàng loạt các cuốn sách và tờ rơi trở thành khả thi. Khi Martin Luther dịch Kinh Thánh Latin sang tiếng Đức thông dụng, dân chúng đột nhiên có nhu cầu biết chữ: để tự đọc Kinh Thánh, và họ sẽ không cần các linh mục để tìm hiểu lời của Chúa. Kinh Thánh đã trở thành một cuốn sách bán chạy nhất. Và một khi biết chữ, mọi người tiếp tục đọc. Một số thậm chí quyết định viết. Trong vòng chưa đầy một vòng đời, dòng thông tin đã thay đổi từ một tia nước nhỏ thành một dòng nước lũ.

Sự thay đổi đáng kể này cũng vun đắp cho các quy tắc mới để chi phối sự bùng nổ thông tin nhờ công nghệ xếp chữ. Khi nhà nước thế tục củng cố quyền lực, nó thiết lập hệ thống kiểm duyệt và cấp giấy phép để kiềm chế và kiểm soát văn bản in ấn. Bản quyền đã được thiết lập nhằm trao cho tác giả những động lực về pháp lý và kinh tế để họ sáng tác.

Sau đó, sự đấu tranh của giới trí thức khiến cho từ thế kỷ XIX, ở ngày càng nhiều quốc gia, tự do ngôn luận đã được biến thành một quyền được bảo đảm trong hiến pháp. Nhưng các quyền này đi kèm với trách nhiệm. Khi những tờ báo cay độc chà đạp quyền riêng tư hoặc vu khống thanh danh, có nhiều quy tắc sẽ bảo vệ người dân và giúp họ khởi kiện tội phỉ báng.

Tuy nhiên, những thay đổi này trong quản lý nhà nước cũng phản ánh một sự chuyển đổi các giá trị nền tảng sâu sắc hơn, cơ bản hơn. Trong cái bóng của Gutenberg, trước tiên chúng ta bắt đầu nhận ra sức mạnh của chữ viết - và cuối cùng là tầm quan trọng của thông tin lan truyền rộng rãi trong toàn xã hội. Sau nhiều thế kỷ trôi qua, chúng ta đã lựa chọn để có nhiều thông tin hơn chứ không phải là ít hơn, và chống lại sự thái quá của nó không phải bằng sự kiểm duyệt mà chủ yếu thông qua các quy tắc hạn chế việc lạm dụng thông tin.

Khi thế giới chuyển dịch về phía dữ liệu lớn, xã hội sẽ trải qua một cuộc chuyển đổi kiến tạo tương tự trong quá khứ. Dữ liệu lớn đã làm thay đổi nhiều khía cạnh của cuộc sống và cách tư duy của chúng ta, buộc chúng ta phải xem xét lại những nguyên tắc cơ bản trong việc khuyến khích sự tăng trưởng và giảm thiểu nguy cơ gây hại của nó. Tuy nhiên, không giống như các vị tiền bối trong và sau cuộc cách mạng in ấn, chúng ta không có nhiều thế kỷ để điều chỉnh, mà có lẽ chỉ có một vài năm.

Những thay đổi đơn giản đối với các quy định hiện hành sẽ không đủ để quản lý trong thời đại dữ-liệu-lớn hoặc hạn chế mặt tối của dữ liệu lớn. Tình hình thực tế đòi hỏi một thay đổi của mô hình. Việc bảo vệ sự riêng tư đòi hỏi người sử dụng dữ-liệu-lớn phải có trách nhiệm cao hơn đối với các hành động của họ. Đồng thời, xã hội sẽ phải xác định lại khái niệm cốt lõi về công lý nhằm đảm bảo quyền tự do của con người để hành động (và do đó chịu trách nhiệm về những hành động này). Cuối cùng, cần có các tổ

chức và các chuyên gia mới để giải thích các thuật toán phức hợp làm nền tảng cho những phát hiện dữ-liệu-lớn, và để bảo vệ cho những người có thể bị dữ liệu lớn gây tổn hại.

Từ sự riêng tư tới trách nhiệm giải trình

Trong nhiều thập kỷ, một nguyên tắc cơ bản của luật riêng tư trên toàn thế giới đã trao quyền kiểm soát cho các cá nhân bằng cách để cho họ quyết định liệu thông tin cá nhân của họ có được xử lý hay không, như thế nào và ai thực hiện. Trong thời đại Internet, lý tưởng đáng khen ngợi này thường biến thành một hệ thống công thức "xin phép và cho phép". Tuy nhiên trong thời đại của dữ liệu lớn, khi nhiều giá trị của dữ liệu nằm trong các ứng dụng thứ cấp có thể chưa được hình dung từ ban đầu, một cơ chế như vậy không còn phù hợp để đảm bảo sự riêng tư.

Chúng ta hình dung một khuôn khổ riêng tư rất khác cho thời đại dữ-liệu-lớn, một khuôn khổ tập trung ít hơn vào sự đồng ý của cá nhân tại thời điểm thu thập thông tin, và nhiều hơn vào việc buộc những người sử dụng dữ liệu phải chịu trách nhiệm về những gì họ làm. Trong một thế giới như vậy, các công ty sẽ chính thức đánh giá một cuộc tái sử dụng dữ liệu cụ thể dựa trên tác động của nó lên các cá nhân có thông tin riêng tư trong đó. Điều này không nhất thiết phải chi tiết một cách phiền hà trong mọi trường hợp, khi luật riêng tư trong tương lai sẽ xác định những nhóm loại rộng của các ứng dụng, bao gồm cả những loại được cho phép mà không có hoặc chỉ có những biện pháp bảo vệ giới hạn, tiêu chuẩn hóa. Với những sáng kiến mang tính rủi ro hơn, các nhà quản lý sẽ thiết lập những quy tắc nền tảng để người dùng dữ liệu có thể đánh giá những nguy hiểm của việc sử dụng và xác định những gì cần tránh hoặc làm giảm thiểu tác hại tiềm ẩn. Điều này khuyến khích việc tái sử dụng sáng tạo của dữ liệu, trong khi đồng thời nó đảm bảo các biện pháp đầy đủ được thực hiện sao cho các cá nhân không bị tổn hại.

Tiến hành đánh giá chính thức việc ứng dụng dữ-liệu-lớn một cách đúng đắn và áp dụng các kết quả của nó một cách chính xác sẽ đem lại những lợi ích hữu hình cho người sử dụng dữ liệu: họ sẽ được tự do theo đuổi những ứng dụng thứ cấp của dữ liệu cá nhân trong nhiều trường hợp mà không cần phải trở lại các cá nhân để có được sự đồng ý rõ ràng của họ. Ngược lại, những sự đánh giá cẩu thả hoặc thực hiện không tốt các biện pháp bảo vệ sẽ đẩy người sử dụng dữ liệu đối mặt với trách nhiệm pháp lý, bị phạt tiền, và thậm chí có thể truy tố hình sự. Trách nhiệm giải trình của người sử dụng dữ liệu chỉ hiệu quả khi có công cụ hỗ trợ.

Để xem điều này có thể xảy ra trong thực tế như thế nào, hãy lấy ví dụ về dữ liệu hóa của dáng điệu trong Chương Năm. Hãy tưởng tượng rằng một công ty bán một dịch vụ chống trộm xe hơi có sử dụng tư thế ngồi của người lái xe như một hình thức kiểm tra an ninh duy nhất. Sau đó, nó tái phân tích thông tin để dự đoán các "trạng thái đáng chú ý", như liệu người lái xe có buồn ngủ, say rượu hoặc tức giận không, để gửi các tín hiệu nhắc nhở tới những người lái xe khác xung quanh nhằm phòng ngừa tai nạn. Theo những quy định bảo mật hiện nay, công ty có thể cho rằng mình cần thực hiện một đợt "xin phép và cho phép" mới, bởi vì trước đây nó chưa được phép sử dụng các thông tin theo cách thức như vậy. Nhưng theo hệ thống trách nhiệm sử dụng dữ liệu, công ty sẽ đánh giá những nguy cơ của ứng dụng chính của dữ liệu, và nếu thấy chúng ở mức tối thiểu thì họ có thể cứ tiến hành với kế hoạch của mình.

Việc chuyển gánh nặng trách nhiệm từ công chúng sang những người sử dụng dữ liệu là hợp lẽ vì nhiều lý do. Họ hiểu nhiều hơn ai hết, và chắc chắn là nhiều hơn so với người tiêu dùng hay nhà quản lý, về việc họ có ý định sử dụng dữ liệu như thế nào. Bằng cách tự tiến hành đánh giá (hoặc thuê chuyên gia để làm điều đó) họ sẽ tránh được vấn đề tiết lộ các chiến lược kinh doanh bí mật cho người ngoài. Có lẽ quan trọng nhất, những người sử

dụng dữ liệu thu được hầu hết lợi ích của các ứng dụng thứ cấp, vì vậy buộc họ chịu trách nhiệm về các hành động của họ và đặt gánh nặng của việc đánh giá này lên vai họ là hoàn toàn hợp lý.

Với một hệ thống như vậy, người sử dụng dữ liệu sẽ không còn bị luật pháp đòi hỏi phải xóa thông tin cá nhân một khi nó đã phục vụ mục đích chính của nó, như hầu hết các luật về quyền riêng tư hiện nay yêu cầu. Đây là một thay đổi quan trọng, bởi vì như chúng ta đã thấy, chỉ bằng cách khai thác giá trị tiềm ẩn của dữ liệu thì những Maury đương thời mới có thể phát triển bằng cách trích xuất giá trị nhiều nhất từ nó cho lợi ích của chính họ - và của xã hội. Người sử dụng dữ liệu sẽ được phép giữ thông tin cá nhân lâu hơn, mặc dù không phải mãi mãi. Xã hội cần cân nhắc cẩn thận những lợi ích từ việc tái sử dụng này, so với các rủi ro vì tiết lộ quá nhiều.

Để đạt được sự cân bằng hợp lý, các nhà điều hành có thể lựa chọn những khung thời gian khác nhau cho việc tái sử dụng, tùy thuộc vào rủi ro vốn có của dữ liệu, cũng như vào các giá trị của những xã hội khác nhau. Một số quốc gia có thể thận trọng hơn những quốc gia khác, cũng giống như một số loại dữ liệu có thể được xem là nhạy cảm hơn những loại khác. Cách tiếp cận này cũng sẽ xua đuổi nỗi ám ảnh về "bộ nhớ vĩnh hằng" - nguy cơ khiến một người không bao giờ có thể thoát khỏi quá khứ bởi vì các hồ sơ kỹ thuật số luôn luôn có thể được bới lên. Các giới hạn thời gian cũng thúc đẩy những người chủ sở hữu dữ liệu phải tận dụng nó trước khi họ mất nó. Điều này đạt được những gì chúng ta tin là một sự cân bằng tốt hơn cho thời đại dữ-liệu-lớn: các công ty được quyền sử dụng dữ liệu cá nhân lâu hơn, nhưng đổi lại họ phải nhận trách nhiệm về việc sử dụng nó cũng như nghĩa vụ phải xóa nó sau một khoảng thời gian nhất định.

Ngoài sự thay đổi quy định từ "cho phép xem thông tin riêng tư" đến "trách nhiệm sử dụng thông tin riêng tư", chúng ta còn hình dung những đổi mới kỹ thuật để giúp bảo vệ sự riêng tư. Một

cách tiếp cận mới ra đời là khái niệm về "quyền riêng tư khác biệt": cố tình làm mờ dữ liệu sao cho việc truy vấn một tập dữ liệu lớn không tiết lộ những kết quả chính xác mà chỉ những kết quả gần đúng. Điều này sẽ gây khó khăn và tốn kém cho việc liên kết những điểm dữ liệu cụ thể với những con người cụ thể. Ví dụ các chuyên gia về chính sách công nghệ lưu ý rằng Facebook dựa trên một hình thức riêng tư khác biệt khi nó báo cáo thông tin về người sử dụng cho các nhà quảng cáo tiềm năng: các con số báo cáo là gần đúng, do đó, chúng không thể giúp tiết lộ danh tính cá nhân. Lệnh tìm kiếm những phụ nữ châu Á ở Atlanta quan tâm đến yoga Ashtanga sẽ cho ra một kết quả kiểu như "khoảng 400", khiến cho việc sử dụng thông tin để hướng đến một người cụ thể là bất khả thi.

Bước chuyển đổi từ sự cho phép mang tính cá nhân sang trách nhiệm của những người sử dụng dữ liệu là một sự thay đổi cơ bản và thiết yếu, cần thiết cho việc quản trị dữ-liệu-lớn hiệu quả. Nhưng nó không phải là thứ duy nhất.



Internet theo dõi chúng ta!

Con người so với dự đoán

Tòa án buộc con người chịu trách nhiệm cho hành động của họ. Khi thẩm phán đưa ra các phán quyết công tâm sau một phiên xét xử công bằng, thì công lý được thực hiện. Tuy nhiên, trong kỷ nguyên của dữ liệu lớn, quan niệm của chúng ta về công lý cần được xác định lại để bảo tồn được ý chí tự do mà với nó con người được lựa chọn các hành động của mình.

Trước thời đại dữ liệu lớn, quyền tự do cơ bản này là rõ ràng tới mức, trong thực tế, nó hầu như không cần phải được nói ra. Xét cho cùng, đó là cách thức hệ thống pháp luật của chúng ta hoạt động: chúng ta buộc con người chịu trách nhiệm về các hành vi của họ bằng cách đánh giá những gì họ đã làm. Ngược lại, với dữ liệu lớn, chúng ta có thể dự đoán các hành động của con người ngày càng chính xác. Điều này cám dỗ chúng ta phán xét con người không phải với những gì họ đã làm, mà với những gì chúng ta dự đoán họ sẽ làm.

Trong thời đại dữ-liệu-lớn chúng ta sẽ phải mở rộng sự hiểu biết của mình về công lý, và đòi hỏi nó bao gồm những biện pháp bảo vệ cho quyền hành động của con người nhiều nhất có thể như chúng ta hiện đang bảo vệ quy trình công bằng. Nếu không có những biện pháp bảo vệ như vậy thì ý tưởng cốt lõi về công lý có thể bị suy yếu hoàn toàn.

Bằng việc bảo đảm quyền hành động của con người, chúng ta đảm bảo rằng phán xét của chính phủ đối với hành vi của chúng ta là được dựa trên những hành động thực tế, chứ không chỉ đơn giản là trên phân tích dữ liệu lớn. Vì vậy, chính phủ chỉ có thể buộc chúng ta chịu trách nhiệm về những hành động quá khứ của chúng ta, chứ không phải những hành động tương lai từ những dự đoán thống kê. Và khi nhà nước phán xét những hành động trước đây, họ phải tránh việc chỉ dựa trên dữ liệu lớn. Ví dụ có chín công ty bị nghi ngờ gian lận giá. Chúng ta hoàn toàn có thể chấp nhận việc sử dụng phân tích dữ-liệu-lớn để xác định việc câu kết với nhau, giúp nhà chức trách điều tra và xây dựng một bản án bằng cách sử dụng những phương tiện truyền thống. Nhưng các công ty này không thể bị buộc tội chỉ vì dữ liệu lớn cho thấy rằng họ có thể phạm tội.

Một nguyên tắc tương tự nên được áp dụng không chỉ đối với cơ quan chính phủ, khi các doanh nghiệp đưa ra những quyết định rất quan trọng về chúng ta - thuê hoặc sa thải, cho vay, hoặc từ

chối một thẻ tín dụng. Khi họ căn cứ các quyết định này chủ yếu trên các dự đoán dữ-liệu-lớn, chúng ta phải thực hiện một số biện pháp bảo vệ. Thứ nhất là tính công khai: công bố dữ liệu và thuật toán làm cơ sở cho dự đoán gây ảnh hưởng đến một cá nhân. Thứ hai là sự chứng nhận: yêu cầu thuật toán được chứng nhận có thể sử dụng cho những mục đích nhạy cảm nhất định, bởi một bên thứ ba có chuyên môn và tính hợp pháp. Thứ ba là sự phản bác: xác định những cách thức cụ thể mà người dân có thể bác bỏ một dự đoán về bản thân họ. (Điều này tương tự với truyền thống trong khoa học về việc tiết lộ mọi yếu tố có thể làm suy yếu các kết quả của một nghiên cứu.)

Quan trọng nhất, một sự đảm bảo về quyền được hành động của con người sẽ chống lại mối đe dọa của một chế độ độc tài của dữ liệu, trong đó chúng ta ban cho dữ liệu nhiều ý nghĩa và tầm quan trọng hơn mức mà nó xứng đáng được nhận.

Một điều cũng không kém phần quan trọng là chúng ta cần bảo vệ trách nhiệm cá nhân. Với rất nhiều dữ liệu dường như khách quan trong tầm tay, người ta dễ có khuynh hướng phi cảm xúc hóa và phi cá nhân hóa quá trình ra quyết định. Người ta sẽ dựa trên các thuật toán thay vì các đánh giá chủ quan, và trình bày các quyết định không bằng ngôn ngữ của trách nhiệm cá nhân mà bằng những rủi ro "khách quan" hơn, cùng với việc phòng tránh chúng.

Ví dụ dữ liệu lớn có thể được dùng để dự đoán những ai có khả năng phạm tội và đặt họ thành đối tượng cần xử lý đặc biệt, rà soát liên tục để giảm rủi ro. Người được phân loại theo cách thức này có thể cảm thấy, và đúng là như thế, rằng họ đang bị trừng phạt nhưng lại chẳng bao giờ được đối mặt và chịu trách nhiệm về hành vi thực tế. Hãy tưởng tượng rằng một thuật toán xác định một thiếu niên nào đó có khả năng rất cao sẽ phạm một trọng tội trong ba năm tới. Kết quả là các nhà chức trách chỉ định một nhân viên xã hội tới thăm cậu ta mỗi tháng một lần, để canh

chừng và cố gắng giúp cậu tránh xa rắc rối.

Nếu thiếu niên đó và người thân, bạn bè, thầy cô giáo, hoặc nơi cậu ta làm việc xem các chuyến thăm như một sự kỳ thị, khi đó sự can thiệp có tác dụng như một hình phạt, cho một hành động chưa hề xảy ra. Và tình hình cũng chẳng tốt hơn bao nhiêu nếu các chuyến thăm được xem như một nỗ lực để làm giảm khả năng của các vấn đề tương lai - như một cách để giảm thiểu rủi ro - chứ không phải một sự trừng phạt. Càng chuyển nhiều từ việc buộc người ta chịu trách nhiệm về những hành vi của mình sang việc giảm thiểu rủi ro trong xã hội nhờ các biện pháp dựa trên phân tích dữ liệu, chúng ta càng làm giảm giá trị của lý tưởng về trách nhiệm cá nhân. Nhà nước mang tính dự báo là nhà nước vú em, và không chỉ có thế. Nếu nhà nước căn cứ nhiều quyết định trên các dự đoán và mong muốn giảm thiểu rủi ro, thì những lựa chọn cá nhân của chúng ta - và do đó tự do cá nhân của chúng ta để hành động - không còn ý nghĩa nữa. Nếu không biết lỗi thì cũng không biết vô tội. Chấp nhận một cách tiếp cận như vậy sẽ không cải thiện xã hội của chúng ta mà làm nó nghèo đi.

Phá vỡ hộp đen

Các hệ thống máy tính hiện nay quyết định dựa trên các nguyên tắc được lập trình một cách rõ ràng. Do đó, khi một quyết định bị sai lệch, điều đôi khi không thể tránh khỏi, chúng ta có thể quay trở lại và tìm ra lý do máy tính đã làm như vậy. Ví dụ chúng ta có thể điều tra những câu hỏi như “Tại sao hệ thống lái tự động lại nghiêng máy bay cao hơn năm độ khi một bộ cảm biến bên ngoài phát hiện sự gia tăng độ ẩm đột ngột?”. Mã máy tính ngày nay có thể được mở ra và kiểm tra, và những ai biết giải thích nó có thể theo dõi và hiểu được cơ sở cho các quyết định của nó, bất kể nó phức tạp ra sao.

Tuy nhiên với phân tích dữ-liệu-lớn, việc truy xuất nguồn gốc

này sẽ khó khăn hơn nhiều. Cơ sở cho các dự đoán của một thuật toán thường quá phức tạp đối với hầu hết mọi người.

Khi máy tính đã được lập trình một cách rõ ràng để làm theo các hướng dẫn, như với chương trình dịch từ tiếng Nga sang tiếng Anh ban đầu của IBM vào năm 1954, một người có thể dễ dàng hiểu tại sao phần mềm lại thay thế một từ bằng một từ khác. Nhưng Google Translate kết hợp hàng tỷ trang dịch vào đánh giá của nó như liệu từ tiếng Anh "light" cần được dịch thành "lumière (ánh sáng)" hay "léger (nhẹ)" trong tiếng Pháp (có nghĩa là liệu từ đó đề cập đến độ sáng hay trọng lượng). Một con người không thể nào lần ra những lý do chính xác cho các lựa chọn từ ngữ của chương trình bởi vì chúng được dựa trên số lượng đồ sộ của dữ liệu và rất nhiều tính toán thống kê.

Dữ liệu lớn hoạt động ở quy mô vượt quá sự hiểu biết thông thường của chúng ta. Ví dụ mối liên hệ Google đã phát hiện giữa một số ít các thuật ngữ tìm kiếm và dịch cúm là kết quả của thử nghiệm 450 triệu mô hình toán học. Ngược lại, Cynthia Rudin ban đầu đã thiết kế 106 dự đoán cho việc liệu một hố ga có thể phát nổ, và cô có thể giải thích cho các nhà quản lý của Con Edison lý do chương trình của cô lại ưu tiên các địa điểm kiểm tra như nó đã làm. Tính chất "có thể giải thích được" là vô cùng quan trọng đối với chúng ta, những người có xu hướng muốn biết tại sao, chứ không chỉ là cái gì. Nhưng điều gì sẽ xảy ra nếu thay vì 106 dự đoán, hệ thống tự động đưa ra con số 601 dự đoán, mà phần lớn trong đó có mức ưu tiên rất thấp, nhưng khi gộp với nhau lại cải thiện độ chính xác của mô hình? Cơ sở cho bất kỳ dự đoán nào cũng có thể vô cùng phức tạp. Vậy cô ấy có thể nói gì với các nhà quản lý để thuyết phục họ tái phân bổ ngân sách hạn chế của họ?

Trong những kịch bản này, chúng ta có thể nhìn thấy rủi ro rằng các dự đoán dữ-liệu-lớn, cùng các thuật toán và các bộ dữ liệu phía sau chúng, sẽ trở thành những hộp đen chẳng hề có trách

nhiệm gì với chúng ta, chẳng có khả năng truy xuất nguồn gốc, chẳng khiến chúng ta tự tin. Để ngăn chặn điều này, dữ liệu lớn sẽ đòi hỏi sự giám sát và minh bạch, mà đến phiên chúng lại đòi hỏi những loại chuyên môn và tổ chức mới. Thời gian gần đây, các chuyên gia về bảo mật máy tính và tính riêng tư đã xuất hiện để xác nhận các công ty đang thực hiện đúng theo các biện pháp tốt nhất được xác lập bởi các cơ quan như Tổ chức Quốc tế về Tiêu chuẩn hóa (ISO) (được thành lập để giải quyết nhu cầu mới về các hướng dẫn trong lĩnh vực này).

Dữ liệu lớn sẽ đòi hỏi một nhóm người mới để đảm nhận vai trò này. Có lẽ họ sẽ được gọi là "các nhà thuật toán". Họ có thể có hai hình thức - thực thể độc lập để giám sát các công ty từ bên ngoài, và nhân viên hoặc các phòng ban để giám sát chúng từ bên trong.

SỰ TRỖI DẬY CỦA NHÀ THUẬT TOÁN

Các nhà chuyên môn mới này sẽ là chuyên gia trong các lĩnh vực khoa học máy tính, toán học, và thống kê; họ sẽ là những người nhận xét các phân tích và dự đoán dữ-liệu-lớn. Các nhà thuật toán sẽ thực hiện một lời thề về công bằng và bảo mật, giống như các nhà kế toán và một số nhà chuyên môn khác hiện nay. Họ sẽ đánh giá việc chọn nguồn dữ liệu, sự lựa chọn các công cụ phân tích và dự báo, bao gồm cả các thuật toán và mô hình, và giải thích kết quả. Trong trường hợp tranh chấp, họ sẽ có quyền truy cập vào các thuật toán, các phương pháp thống kê, và các bộ dữ liệu dùng để đưa ra một quyết định cụ thể.

Nếu như có một nhà thuật toán tại Bộ An ninh Nội địa vào năm 2004, ông đã có thể ngăn chặn được việc cơ quan này tạo ra một danh sách cấm bay sai lầm đến nỗi bao gồm cả Thượng nghị sĩ Kennedy. Những trường hợp gần đây hơn mà các nhà thuật toán có lẽ giúp ích được đã xảy ra ở Nhật Bản, Pháp, Đức, và Ý. Người ta phàn nàn rằng tính năng "tự động hoàn chỉnh" của Google đã

phỉ báng họ bằng việc tạo ra một danh sách các thuật ngữ tìm kiếm phổ biến gắn với tên người gõ vào. Danh sách này chủ yếu dựa vào tần số của lệnh các tìm kiếm trước đây: các thuật ngữ được xếp hạng theo xác suất toán học của chúng. Tuy nhiên, trong chúng ta, ai mà lại không sôi máu nếu từ "tội phạm" hay "gái điếm" xuất hiện bên cạnh tên của chúng ta khi những đối tác tiềm năng hay người yêu lên mạng tìm thông tin về chúng ta?

Có thể hình dung rằng các nhà thuật toán, khi mang đến một phương pháp tiếp cận theo định hướng thị trường đối với các vấn đề như thế này, có thể tránh được những hình thức phiền phức hơn liên quan đến luật pháp. Họ sẽ thỏa mãn một nhu cầu tương tự như nhu cầu mà các nhà kế toán và kiểm toán đã đáp ứng khi xuất hiện trong những năm đầu của thế kỷ XX để xử lý tình trạng tràn ngập thông tin tài chính. Bằng cách cung cấp dịch vụ giám sát tài chính, thành phần chuyên gia mới xuất hiện này đã củng cố niềm tin của xã hội vào nền kinh tế. Dữ liệu lớn có thể và cần được hưởng lợi từ việc tăng cường niềm tin tương tự mà các nhà thuật toán sẽ cung cấp.

CÁC NHÀ THUẬT TOÁN BÊN NGOÀI

Chúng ta hình dung rằng các nhà thuật toán bên ngoài sẽ đóng vai trò như kiểm toán viên độc lập để xem xét tính chính xác hay hiệu lực của các dự đoán dữ-liệu-lớn, bất cứ khi nào chính phủ yêu cầu, chẳng hạn như theo lệnh của tòa án. Họ cũng có thể nhận các công ty dữ-liệu-lớn làm khách hàng, thực hiện "kiểm toán" cho các công ty muốn có sự hỗ trợ chuyên môn. Và họ có thể xác nhận tính đúng đắn của các ứng dụng dữ-liệu-lớn như các kỹ thuật chống gian lận hoặc các hệ thống kinh doanh chứng khoán. Cuối cùng, các nhà thuật toán bên ngoài được chuẩn bị để tư vấn cho các cơ quan chính phủ về cách tốt nhất để sử dụng dữ liệu lớn trong khu vực công. Cũng như trong y học, pháp luật, và các ngành nghề khác, chúng ta hình dung nghề nghiệp mới này

có quy định riêng của nó với một bộ quy tắc ứng xử. Tính vô tư, bảo mật, năng lực, và tính chuyên nghiệp của các nhà thuật toán được thực thi bởi các quy tắc trách nhiệm chặt chẽ; nếu không tuân thủ những tiêu chuẩn này, họ sẽ phải đối mặt với pháp luật. Họ cũng có thể được yêu cầu phục vụ như nhân chứng chuyên môn trong các phiên tòa, hoặc hoạt động như “các chủ tọa”, các chuyên gia được bổ nhiệm bởi các thẩm phán để hỗ trợ về các vấn đề kỹ thuật trong những vụ án đặc biệt phức tạp.

Hơn nữa, những người tin rằng họ đã bị tổn hại bởi các dự đoán dữ-liệu-lớn - một bệnh nhân bị từ chối phẫu thuật, một tù nhân bị từ chối tạm tha, một người bị từ chối cho vay thế chấp - có thể tìm đến các nhà thuật toán cũng giống như họ đã tìm đến các luật sư để được giúp đỡ trong việc tìm hiểu và phản đối những quyết định đó.

CÁC NHÀ THUẬT TOÁN NỘI BỘ

Các nhà thuật toán nội bộ làm việc trong một tổ chức để giám sát các hoạt động dữ liệu lớn của nó. Họ quan tâm không chỉ tới lợi ích của công ty mà còn tới lợi ích của những người bị ảnh hưởng bởi các phân tích dữ-liệu-lớn của nó. Họ giám sát các hoạt động dữ-liệu-lớn, và họ là những điểm liên lạc đầu tiên cho bất cứ ai cảm thấy bị tổn hại do những dự đoán dữ liệu lớn của tổ chức. Họ cũng điều chỉnh các phân tích dữ-liệu-lớn về tính toàn vẹn và chính xác trước khi cho phép công bố chúng. Để thực hiện vai trò thứ nhất kể trên, các nhà thuật toán phải có được một mức độ tự do và khách quan nhất định trong tổ chức mà họ làm việc.

Ý niệm về một người làm việc cho một công ty nhưng lại khách quan đối với các hoạt động của nó có vẻ lạ đời, nhưng những tình huống như vậy thực sự khá phổ biến. Các bộ phận giám sát tại những tổ chức tài chính lớn là một ví dụ, như các hội đồng quản trị tại nhiều công ty, có trách nhiệm với các cổ đông, chứ không phải với ban quản lý. Và nhiều công ty truyền thông, trong đó có

New York Times và *Washington Post*, sử dụng các thanh tra có trách nhiệm chính là để bảo vệ niềm tin của công chúng. Những nhân viên này xử lý các khiếu nại của độc giả và thường trừng phạt công ty của họ một cách công khai khi họ xác định nó đã làm sai.

Và có một hình thức còn gần hơn nữa với nhà thuật toán nội bộ - một chuyên gia chịu trách nhiệm đảm bảo thông tin cá nhân không bị lạm dụng trong thiết chế của công ty. Ví dụ Đức yêu cầu các công ty với quy mô nhất định (thường có mười nhân viên hoặc nhiều hơn tham gia vào việc xử lý thông tin cá nhân) phải chỉ định một đại diện bảo vệ dữ liệu. Từ những năm 1970, những đại diện nội bộ này đã xây dựng một hệ thống đạo đức nghề nghiệp và một tinh thần đồng đội. Họ thường xuyên gặp gỡ để chia sẻ kinh nghiệm, đào tạo, và có phương tiện truyền thông cùng các hội thảo chuyên ngành riêng. Hơn nữa, họ đã thành công trong việc duy trì bổn phận kép đối với tổ chức của mình *và* đối với nghĩa vụ của họ là những người nhận xét khách quan, để hoạt động được như những thanh tra bảo vệ dữ liệu trong khi vẫn mang đến những giá trị về sự riêng tư của thông tin trong các hoạt động của công ty. Chúng ta tin rằng những nhà thuật toán nội bộ có thể làm cùng điều như vậy.

Quản lý các ông trùm dữ liệu

Dữ liệu đối với xã hội thông tin giống như nhiên liệu đối với nền kinh tế công nghiệp: thứ tài nguyên quan trọng tạo năng lượng cho những đổi mới mà con người dựa vào. Nếu không có một nguồn cung cấp dữ liệu phong phú, sôi động và một thị trường mạnh mẽ cho các dịch vụ thì sự sáng tạo và hiệu suất tiềm năng có thể bị kiềm hãm.

Trong chương này chúng ta đã đặt ra ba chiến lược mới cơ bản cho quản lý dữ-liệu-lớn, liên quan đến sự riêng tư, xu hướng, và kiểm tra theo thuật toán. Chúng ta tự tin rằng với những chiến

lược này, mặt tối của dữ liệu lớn sẽ được khống chế. Tuy nhiên, khi ngành công nghiệp dữ liệu lớn mới mẻ phát triển, một thách thức quan trọng nữa sẽ là bảo vệ các thị trường dữ liệu lớn cạnh tranh. Chúng ta phải ngăn chặn sự nổi lên của các ông trùm dữ liệu thế-kỷ-hai-mươi-mốt, cũng giống như các ông trùm tư bản của thế kỷ XIX đã thống trị ngành đường sắt, sản xuất thép, và mạng lưới điện báo của Mỹ.

Để kiểm soát các nhà công nghiệp trước đây, Hoa Kỳ đã thiết lập các quy định chống độc quyền cực kỳ linh hoạt. Ban đầu được áp dụng cho các tuyến đường sắt trong những năm 1800, về sau chúng được áp dụng cho các công ty "gác cổng" cho dòng chảy thông tin mà các doanh nghiệp khác phụ thuộc vào, từ National Cash Register trong những năm 1910, đến IBM trong những năm 1960, Xerox trong những năm 1970, AT&T trong những năm 1980, Microsoft trong những năm 1990, và Google ngày nay. Công nghệ mà các công ty này mở đường đã trở thành những phần cốt lõi của "cơ sở hạ tầng thông tin" của nền kinh tế, và đòi hỏi sức mạnh của pháp luật để ngăn chặn sự thống trị không lành mạnh.

Để đảm bảo các điều kiện cho một thị trường năng động cho dữ liệu lớn, chúng ta sẽ cần các biện pháp tương tự với những biện pháp đã thiết lập sự cạnh tranh và giám sát trong các lĩnh vực công nghệ trước đây. Chúng ta cần cho phép giao dịch dữ liệu, chẳng hạn như thông qua cấp phép và khả năng tương tác. Điều này đặt ra vấn đề liệu xã hội có thể được hưởng lợi từ một "quyền độc quyền" về dữ liệu được thiết lập một cách cẩn trọng và cân bằng (tương tự như quyền sở hữu trí tuệ). Phải thừa nhận rằng đạt được điều này sẽ là một thách thức lớn đối với các nhà hoạch định chính sách - và một thứ đầy rủi ro đối với phần còn lại của chúng ta.

Rõ ràng không thể nói trước một công nghệ sẽ phát triển như thế nào; thậm chí dữ liệu lớn cũng không thể dự đoán bản thân nó sẽ

tiến triển ra sao. Các cơ quan quản lý phải có sự cân bằng giữa việc hành động một cách thận trọng và mạnh dạn - và lịch sử của luật chống độc quyền cho thấy một phương cách khả thi.

Luật chống độc quyền kiềm chế sự lạm dụng sức mạnh. Tuy nhiên, điều đáng lưu ý là các nguyên tắc của nó được dịch chuyển rất trơn tru từ lĩnh vực này sang lĩnh vực khác, và xuyên suốt các loại hình khác nhau của các ngành công nghiệp mạng. Nó đúng là loại luật định vững chãi - không làm lợi cho một loại công nghệ hơn một loại khác - rất hữu ích, vì nó bảo vệ cạnh tranh mà không cần phỏng chừng để làm nhiều hơn thế. Do đó, việc chống độc quyền có thể giúp dữ liệu lớn tiến lên phía trước giống như nó đã làm đối với các tuyến đường sắt. Ngoài ra, với vai trò thuộc trong số những chủ sở hữu dữ liệu lớn nhất thế giới, các chính phủ phải phát hành dữ liệu của họ một cách công khai. Điều đáng khích lệ là một số chính phủ đã làm những điều này - ít ra ở một mức độ nào đó.

Bài học về quy định chống độc quyền là một khi các nguyên tắc bao quát đã được xác định, các cơ quan quản lý có thể thực thi chúng để đảm bảo mức độ bảo vệ và hỗ trợ cần thiết. Tương tự như vậy, ba chiến lược chúng ta đã đưa ra - chuyển sự bảo vệ quyền riêng tư từ hình thức cho phép của cá nhân sang trách nhiệm của người sử dụng dữ liệu, gìn giữ quyền hành động của con người trong bối cảnh dự đoán, và thiết lập nhóm nghề mới gồm các “kiểm toán viên” dữ-liệu-lớn mà chúng ta gọi các nhà thuật toán - có thể đóng vai trò như nền tảng cho sự quản trị hiệu quả và công bằng về thông tin trong thời đại dữ-liệu-lớn.

Trong nhiều lĩnh vực, từ công nghệ hạt nhân tới công nghệ sinh học, chúng ta xây dựng các công cụ và rồi phát hiện chúng có thể làm hại mình. Chỉ đến lúc đó chúng ta mới đưa ra các cơ chế an toàn để bảo vệ mình trước những công cụ như thế. Về phương diện này, dữ liệu lớn cũng song hành cùng các lĩnh vực khác của xã hội và đưa ra những thách thức mà không có các giải pháp

tuyệt đối, chỉ có những câu hỏi liên tiếp về cách chúng ta sắp đặt thế giới của mình. Mỗi thế hệ lại phải giải quyết những vấn đề này một lần nữa. Nhiệm vụ của chúng ta là đánh giá các mối nguy của công nghệ mạnh mẽ này, hỗ trợ sự phát triển của nó - và hưởng thụ những phần thưởng của nó.

Giống như việc in ấn đã dẫn đến những thay đổi trong cách xã hội điều chỉnh chính nó, dữ liệu lớn cũng sẽ làm như vậy. Nó buộc chúng ta phải đối đầu những thách thức mới với những giải pháp mới. Để đảm bảo con người được bảo vệ đồng thời với việc công nghệ được đẩy mạnh, chúng ta không thể để cho dữ liệu lớn phát triển vượt ngoài tầm khả năng của con người để định hình công nghệ này.

10. TIẾP THEO

MIKE FLOWERS LÀ MỘT LUẬT SƯ ở văn phòng chưởng lý hạt Manhattan trong những năm 2000, khởi tố tất cả mọi thứ, từ những vụ giết người cho tới những tội phạm ở Wall Street. Sau đó ông chuyển sang một công ty luật doanh nghiệp sang trọng. Qua một năm nhàm chán sau bàn làm việc, ông đã quyết định bỏ công việc này. Vì muốn tìm kiếm một cái gì đó có ý nghĩa hơn, ông nghĩ tới việc giúp đỡ xây dựng lại Iraq. Một đối tác thân thiết của Flowers tại công ty đã gọi điện cho một số nhân vật cấp cao. Và thế là ông cũng không ngờ mình nhanh chóng chuyển đến Vùng Xanh, khu vực an toàn cho quân đội Mỹ trong trung tâm Baghdad, tham gia đội ngũ pháp lý cho tòa án xử Saddam Hussein.

Hầu hết các công việc của ông hóa ra là về hậu cần chứ không phải pháp lý. Ông phải xác định những khu vực tình nghi là những ngôi mộ tập thể để cử các nhà điều tra tới. Ông phải đưa các nhân chứng vào Vùng Xanh mà không khiến họ sa vào những vụ nổ vì các loại bom tự chế (IED), một thực tế nghiệt ngã hàng ngày. Ông nhận thấy rằng quân đội xem các nhiệm vụ này như các bài toán thông tin. Và dữ liệu sẽ là cứu cánh. Các nhà phân tích tình báo sẽ kết hợp các báo cáo thực địa với những chi tiết về địa điểm, thời gian, và thương vong của các cuộc tấn công IED trong quá khứ để dự đoán tuyến đường an toàn nhất cho ngày hôm đó.

Khi trở về thành phố New York một vài năm sau đó, Flowers nhận ra rằng những phương pháp này cho thấy một cách thức chống lại tội phạm tốt hơn so với những gì ông từng có lúc còn là một công tố viên. Và ông đã tìm thấy một người đồng cảm thật sự, thị trưởng của thành phố, Michael Bloomberg, người đã tạo nên thời vận của mình từ dữ liệu bằng cách cung cấp thông tin tài chính cho các ngân hàng. Flowers được đưa vào một đơn vị đặc biệt, với nhiệm vụ xử lý số liệu để có thể vạch mặt những kẻ

lừa đảo trong cuộc khủng hoảng thế chấp dưới chuẩn trong năm 2009. Đơn vị đã thành công tới mức một năm sau đó thị trưởng Bloomberg đã đề nghị mở rộng phạm vi của nó. Flowers đã trở thành "giám đốc phân tích" đầu tiên của thành phố. Nhiệm vụ của ông: xây dựng một đội ngũ các nhà khoa học dữ liệu tốt nhất ông có thể tìm thấy và khai thác những kho tàng thông tin chưa được khám phá của thành phố nhằm gặt hái hiệu quả trong mọi lĩnh vực.

Flowers xem xét mạng lưới quen biết rộng lớn của mình để tìm đúng người. "Tôi không quan tâm tới những nhà thống kê rất giàu kinh nghiệm", ông nói. "Tôi có chút lo ngại rằng họ sẽ phải miễn cưỡng chấp nhận cách tiếp cận mới này để giải quyết vấn đề". Trước đó, khi ông phỏng vấn những chuyên gia thống kê truyền thống cho dự án gian lận tài chính, họ có xu hướng nêu những lo ngại nhà nghề về các phương pháp toán học. "Tôi thậm chí không nghĩ về mô hình mà tôi sẽ sử dụng. Tôi muốn sự hiểu biết sâu sắc để hành động, và đó là tất cả những gì tôi quan tâm", ông nói. Cuối cùng thì ông đã chọn một nhóm năm người mà ông gọi là "những đứa trẻ". Tất cả, trừ một người, đều trong những chuyên ngành kinh tế, mới chỉ tốt nghiệp một hoặc hai năm và không có nhiều kinh nghiệm sống ở một thành phố lớn, và tất cả đều thể hiện tố chất sáng tạo.

Một trong số những thách thức đầu tiên mà nhóm đã giải quyết là "chuyển đổi bất hợp pháp" - thật ra nghĩa là đem chia một chỗ cư trú thành nhiều đơn vị nhỏ hơn để có thể chứa được tới mười lần nhiều hơn số lượng người mà nó đã được thiết kế để chứa. Có nhiều nguy cơ về hỏa hoạn, cũng như về chứa chấp tội phạm, ma túy, bệnh tật, và sâu bệnh. Một mớ dây cáp điện có thể chạy ngoằn ngoèo trong các bức tường, còn những tấm sưởi bị bỏ một cách nguy hiểm trên những khăn trải giường. Những người sống chật trội như vậy thường xuyên bị chết trong các đám cháy. Năm 2005 hai nhân viên cứu hỏa đã thiệt mạng khi cố gắng giải cứu người dân. Thành phố New York có khoảng 25.000 khiếu nại

về chuyển đổi bất hợp pháp mỗi năm, nhưng chỉ có 200 thanh tra để xử lý chúng. Dường như không có cách nào tốt để loại ra những trường hợp chỉ đơn giản là phiền hà từ những người dễ bốc hỏa. Tuy nhiên đối với Flowers và những đứa trẻ của ông, điều này trông giống như một bài toán có thể được giải với rất nhiều dữ liệu.

Họ bắt đầu với một danh sách của các bất động sản trong thành phố - tất cả có 900.000. Tiếp theo, họ bổ sung dữ liệu từ 19 cơ quan khác nhau cho biết các thông tin như liệu chủ sở hữu tòa nhà có vi phạm quy định trả tiền thuế bất động sản, từng có thủ tục tố tụng tịch thu nhà, có những bất thường trong việc sử dụng các dịch vụ tiện ích, hoặc bị cắt dịch vụ vì không thanh toán. Họ cũng đưa vào thông tin về loại hình của tòa nhà và khi nào nó được xây dựng, cộng với những lần gọi xe cứu thương, tỷ lệ tội phạm, khiếu nại... Sau đó, họ so sánh tất cả các thông tin này đối với năm năm của dữ liệu về các vụ cháy được xếp hạng theo mức độ nghiêm trọng và tìm kiếm những mối tương quan để tạo ra một hệ thống có thể dự đoán được những khiếu nại nào phải được điều tra khẩn cấp nhất.

Ban đầu, phần lớn các dữ liệu không phải ở dưới hình thức có thể sử dụng được. Ví dụ các cơ quan lưu trữ của thành phố đã không sử dụng một cách thức tiêu chuẩn đơn nhất để mô tả vị trí, mỗi cơ quan và bộ phận dường như có cách tiếp cận riêng của mình. Sở công trình gán mỗi cấu trúc với một số nhà duy nhất. Sở bảo quản nhà ở có một hệ thống đánh số khác. Sở thuế cung cấp cho mỗi bất động sản một định danh dựa theo quận, khu phố và mảnh đất. Cảnh sát sử dụng tọa độ Descartes. Sở cứu hỏa lại trông cậy vào một hệ thống khoảng cách tới các “hộp gọi” tương ứng với vị trí của các trạm cứu hỏa, mặc dù những hộp gọi không còn tồn tại nữa. “Những đứa trẻ” của Flowers đã đối mặt với sự hỗn độn này bằng cách đặt ra một hệ thống định danh các tòa nhà theo phương thức sử dụng một khu vực nhỏ ở phía trước bất động sản dựa trên tọa độ Descartes, và sau đó rút ra dữ liệu vị trí

từ những cơ sở dữ liệu của các cơ quan khác. Phương pháp của họ vốn không chính xác, nhưng số lượng lớn các dữ liệu họ có thể sử dụng đã bù đắp lại cho những khiếm khuyết này.

Tuy nhiên các thành viên trong nhóm đã không thỏa mãn khi chỉ xử lý những con số. Họ đã đi thực địa để xem những thanh tra làm việc. Họ đã ghi chép rất nhiều và hỏi các chuyên gia về mọi thứ. Khi một chỉ huy tóc hoa râm lầm bầm rằng tòa nhà họ sắp kiểm tra không có vấn đề gì đâu, các thành viên trong nhóm nghiên cứu đều hỏi tại sao ông lại cảm thấy chắc chắn như vậy. Ông ta có thể không hoàn toàn nói ra, nhưng "những đứa trẻ" dần xác định được rằng trực giác của ông ta là dựa trên những viên gạch mới ở ngoại thất của tòa nhà, điều khiến ông ta nghĩ rằng người chủ sở hữu quan tâm đến nơi này.

"Những đứa trẻ" trở lại nơi làm việc và tự hỏi làm thế nào để đưa yếu tố "phần gạch mới xây" vào mô hình của họ như một tín hiệu. Xét cho cùng, những viên gạch đâu có được dữ liệu hóa - đúng ra là chưa! Nhưng chắc chắn rằng bất kỳ phần xây mới bên ngoài nào cũng phải có giấy phép của thành phố để thực hiện. Thế là việc thêm thông tin giấy phép đã cải thiện hiệu suất dự đoán của hệ thống, khi cho thấy một số tòa nhà trong diện "đáng ngờ" có lẽ không mang những rủi ro lớn.

Có những lúc các phân tích cũng chỉ ra rằng một số cách thức làm việc lâu đời không phải là cách tốt nhất, cũng giống như các tuyển trạch viên trong *Moneyball* phải chấp nhận những thiếu sót trực giác của họ. Ví dụ số cuộc gọi đến đường dây nóng khiếu nại "311" của thành phố vốn được xem là chỉ báo về những tòa nhà nào cần sự chú ý nhất. Nhiều cuộc gọi hơn tương ứng với những vấn đề nghiêm trọng hơn. Nhưng điều này hóa ra là một sự nhầm lẫn. Một con chuột được phát hiện ở khu Đông Bắc sang trọng có thể gây ra đến ba mươi cuộc gọi trong vòng một giờ, nhưng có thể cần cả một tiểu đoàn động vật gặm nhấm mới làm cho những người dân ở Bronx cảm thấy cần bấm số 311. Tương

tự như vậy, phần lớn những khiếu nại về một sự chuyển đổi bất hợp pháp nào đó có thể là về tiếng ồn, chứ không phải về những điều kiện nguy hiểm.

Tháng 6 năm 2011, Flowers và những đứa trẻ của ông bật công tắc hệ thống của họ. Các khiếu nại thuộc thể loại chuyển đổi bất hợp pháp đã được xử lý hàng tuần. Họ tập hợp những khiếu nại được xếp hạng trong tốp 5 phần trăm về nguy cơ cháy và chuyển chúng tới các thanh tra để theo dõi ngay lập tức. Khi có kết quả trở lại, tất cả mọi người đều sững sờ.

Trước khi có phân tích dữ-liệu-lớn, các thanh tra viên theo dõi các khiếu nại mà họ xem là nghiêm trọng nhất, nhưng họ chỉ nhận thấy tình trạng nghiêm trọng đủ để đưa trát di dời trong 13 phần trăm trường hợp. Bây giờ họ phải phát lệnh đó cho hơn 70 phần trăm các tòa nhà mà họ kiểm tra. Bằng cách chỉ ra những tòa nhà cần sự chú ý của họ nhất, dữ liệu lớn đã cải thiện hiệu quả gấp năm lần. Và công việc này đã khiến các thanh tra viên hài lòng hơn: họ được tập trung vào những vấn đề lớn nhất. Hiệu quả đối với các thanh tra viên cũng có những lợi ích lan tỏa. Những vụ cháy trong các tòa nhà bị chuyển đổi bất hợp pháp có nguy cơ dẫn đến thương tích hoặc tử vong cho nhân viên cứu hỏa 15 lần nhiều hơn so với những vụ cháy khác, vì vậy các sở cứu hỏa cũng mê mẩn nghiên cứu này. Flowers và những đứa trẻ của ông như những thầy pháp với một quả cầu pha lê cho phép họ nhìn thấy tương lai và dự đoán những nơi nào rủi ro nhất. Họ đã lấy những lượng lớn dữ liệu nằm rải rác trong nhiều năm qua, phần lớn không được sử dụng sau khi thu thập, và khai thác nó theo một cách thức mới mẻ để thu được giá trị thực sự. Việc sử dụng một lượng lớn thông tin cho phép họ phát hiện những mối liên hệ bị che giấu trong những lượng nhỏ thông tin hơn. Đó là bản chất của dữ liệu lớn.

Kinh nghiệm của các nhà giả kim thuật phân tích của thành New York làm nổi bật nhiều chủ đề của cuốn sách này. Họ đã sử

dụng một lượng khổng lồ của dữ liệu, và danh sách các tòa nhà trong thành phố đã thể hiện đúng tiêu chí N = tất cả. Dữ liệu này hỗn độn, chẳng hạn thông tin vị trí hoặc hồ sơ xe cứu thương, nhưng điều đó không cản trở được họ. Thật ra, những lợi ích của việc sử dụng nhiều dữ liệu đã vượt hẳn những hạn chế của việc dùng ít thông tin như trước đây. Họ đã có thể đạt được những thành tựu, bởi vì rất nhiều đặc tính của thành phố đã được dữ liệu hóa (tuy không phải một cách nhất quán), cho phép họ xử lý được thông tin.

Những nghi ngờ của các chuyên gia đã phải lùi một bước đối với phương pháp tiếp cận theo định hướng dữ liệu. Đồng thời, Flowers và "những đứa trẻ" của ông đã tiếp tục thử nghiệm hệ thống của họ với các thanh tra viên kỳ cựu, dựa trên kinh nghiệm của họ để làm cho hệ thống hoạt động tốt hơn. Tuy nhiên, lý do quan trọng nhất cho sự thành công của chương trình là nó được thực hiện với sự tin cậy phụ thuộc vào mối tương quan thay vì quan hệ nhân quả.

"Tôi không quan tâm đến nguyên nhân trừ khi nó nói đến hành động", Flowers giải thích. "Nhân quả là cho người khác, và thẳng thắn mà nói, có rất nhiều rủi ro khi bạn bắt đầu nói về quan hệ nhân quả. Tôi không nghĩ rằng có bất kỳ liên hệ nhân quả nào giữa ngày mà một người tiến hành thủ tục tịch biên đối với một bất động sản và việc liệu có khả năng nơi này tiềm ẩn nguy cơ về hỏa hoạn. Tôi cho rằng tư duy như thế là u mê. Và không ai có thể thực sự đứng ra và phát biểu như vậy. Họ sẽ bảo *không, ẩn sâu bên dưới nó vẫn là thế mà*. Nhưng tôi chẳng muốn xuống đến tận đó. Tôi cần một điểm dữ liệu cụ thể để truy cập tới, và cho tôi biết ý nghĩa của nó. Nếu nó quan trọng, chúng tôi sẽ hành động dựa trên đó. Nếu không, chúng tôi sẽ bỏ qua. Anh biết đấy, chúng tôi có những vấn đề hiển hiện, cần được giải quyết. Thực tình vào lúc này, chúng tôi không thể cứ luẩn quẩn, suy nghĩ về những thứ như quan hệ nhân quả được".

Khi dữ liệu nói

Những ảnh hưởng của dữ liệu lớn là khá lớn trên thực tế, khi công nghệ này được áp dụng để tìm lời giải cho các vấn đề thường ngày gây nhiều tranh cãi. Nhưng đó mới chỉ là khởi đầu. Dữ liệu lớn đã sẵn sàng để định hình lại cách chúng ta sống, làm việc, và tư duy. Sự thay đổi chúng ta phải đối mặt, theo một số khía cạnh, thậm chí còn lớn hơn so với những thay đổi mà trước đây đã mở rộng đáng kể phạm vi và quy mô của thông tin trong xã hội. Mặt đất dưới chân chúng ta đang chuyển đổi. Những điều trước đây còn chắc chắn thì lúc này đang bị chất vấn. Dữ liệu lớn đòi hỏi sự tranh luận mới mẻ về bản chất của việc ra quyết định, số phận, công lý. Một thế giới quan mà chúng ta cho rằng được tạo nên từ quan hệ nhân quả đang bị thách thức bởi ưu thế của các mối tương quan. Việc nắm bắt kiến thức, mà trước đây có nghĩa là sự hiểu biết về quá khứ, đang dần chuyển thành khả năng dự đoán tương lai.

Những vấn đề này quan trọng hơn rất nhiều so với những vấn đề nảy sinh khi chúng ta chuẩn bị để khai thác thương mại điện tử, sống với Internet, bước vào thời đại máy tính, hay bỏ đi bàn tính. Ý tưởng cho rằng chúng ta quá quan tâm đến hành trình đi tìm kiếm nguyên nhân - mà trong nhiều trường hợp có thể sẽ thuận lợi hơn nếu tránh né câu hỏi *tại sao* để chuyển sang *cái gì* - cho thấy những vấn đề này là nền tảng cho xã hội và sự tồn tại của chúng ta. Những thách thức do dữ liệu lớn đặt ra có thể cũng không tìm được lời đáp. Thay vào đó, chúng là một phần của cuộc tranh luận vô tận về vị trí của con người trong vũ trụ và cuộc tìm kiếm của con người về ý nghĩa cuộc sống, giữa một thế giới náo nhiệt, hỗn loạn, không thể hiểu nổi.

Xét cho cùng, dữ liệu lớn đánh dấu thời khắc "xã hội thông tin" đã hoàn thành viễn cảnh bao hàm trong tên gọi của nó. Dữ liệu chiếm lĩnh vũ đài trung tâm. Tất cả những bit kỹ thuật số mà chúng ta thu thập bây giờ có thể được khai thác theo những cách

thức mới để phục vụ những mục đích mới và mở khóa cho các dạng giá trị mới. Nhưng điều này đòi hỏi phải có một cách tư duy mới, sẽ thách thức các thể chế của chúng ta và thậm chí cả ý thức của chúng ta về bản sắc. Một điều chắc chắn là lượng dữ liệu sẽ tiếp tục tăng, cũng như sức mạnh để xử lý dữ liệu. Nhưng trong khi hầu hết mọi người đều xem dữ liệu lớn như một vấn đề công nghệ, tập trung vào phần cứng hay phần mềm, chúng tôi lại tin rằng nên hướng sự chú ý sang những điều sẽ xảy ra khi dữ liệu "nói".

Chúng ta có thể thu thập và phân tích nhiều thông tin hơn bao giờ hết. Sự khan hiếm của dữ liệu không còn là đặc tính xác định những nỗ lực của chúng ta để giải thích thế giới. Chúng ta có thể khai thác dữ liệu ở quy mô rộng lớn hơn rất nhiều, và trong một số trường hợp có thể đến gần được với tất cả dữ liệu. Nhưng làm như vậy khiến chúng ta phải hoạt động theo những cách phi truyền thống, và đặc biệt nó sẽ thay đổi suy nghĩ của chúng ta về những gì cấu thành thông tin hữu ích.

Thay vì bị ám ảnh về tính chính xác, tính đúng đắn, sạch sẽ, và tính chắc chắn của dữ liệu, chúng ta có thể để cho một số hạt sạn xen vào. Chúng ta không nên chấp nhận một tập hợp dữ liệu hoàn toàn sai hoặc đúng, nhưng có thể chấp nhận sự hỗn độn để đổi lại việc thu về một tập dữ liệu toàn diện hơn rất nhiều. Thật ra, trong một số trường hợp sự to lớn và hỗn độn thậm chí có thể có lợi, bởi vì khi cố gắng sử dụng chỉ một phần nhỏ và chính xác của dữ liệu, chúng ta cuối cùng đã thất bại trong việc nắm bắt được chiều rộng của chi tiết nơi có chứa rất nhiều kiến thức.

Vì các mối tương quan có thể được tìm thấy nhanh hơn và rẻ hơn so với quan hệ nhân quả, chúng thường thích hợp hơn. Chúng ta sẽ vẫn cần các nghiên cứu nhân quả và các thí nghiệm có kiểm soát với các dữ liệu được giám tuyển cẩn thận trong một số trường hợp, chẳng hạn như thiết kế một chi tiết máy bay quan trọng. Nhưng đối với nhiều nhu cầu hàng ngày, việc biết *cái gì*

chứ không phải *tại sao* là đủ tốt rồi. Và các mối tương quan dữ-liệu-lớn có thể chỉ ra con đường hướng tới các lĩnh vực đầy triển vọng, mà trong đó con người có thể khám phá những mối quan hệ nhân quả.

Các mối tương quan nhanh chóng này cho phép chúng ta tiết kiệm tiền vé máy bay, dự báo dịch cúm, và biết được hố ga hoặc các tòa nhà quá đông đúc nào cần phải kiểm tra trong một thế giới khá hạn hẹp về nguồn lực. Chúng có thể giúp các công ty bảo hiểm y tế cung cấp dịch vụ bảo hiểm mà không cần một kỳ khám sức khỏe, và giảm chi phí nhắc nhở người bệnh dùng thuốc. Ngôn ngữ sẽ được dịch và những chiếc xe sẽ tự lái trên cơ sở các dự đoán được thực hiện thông qua các mối tương quan dữ-liệu-lớn. Walmart có thể biết những hương vị Pop-Tarts nào nên bày ở phía trước cửa hàng trước một cơn bão. (Câu trả lời: dâu đất). Tất nhiên, quan hệ nhân quả là tốt khi bạn nắm bắt được nó. Vấn đề nằm ở chỗ điều này rất khó, và nếu tưởng rằng mình đã tìm thấy nó thì thường chúng ta chỉ tự lừa dối mà thôi.

Những công cụ mới, từ các bộ vi xử lý nhanh hơn và bộ nhớ lớn hơn tới phần mềm và các thuật toán thông minh hơn, chỉ là một phần của lý do chúng ta có thể làm được tất cả những điều này. Dù các công cụ có vai trò quan trọng, một lý do cơ bản hơn là chúng ta có nhiều dữ liệu hơn, bởi vì nhiều khía cạnh của thế giới đang được dữ liệu hóa. Khá chắc chắn rằng tham vọng của con người về việc định lượng thế giới đã có từ rất lâu trước cuộc cách mạng máy tính. Nhưng các công cụ kỹ thuật số tạo thuận lợi cho việc dữ liệu hóa rất nhiều. Điện thoại di động không chỉ theo dõi người mà chúng ta gọi và nơi chúng ta đến, mà các dữ liệu chúng thu thập còn có thể được sử dụng để phát hiện xem chúng ta có đang bị bệnh không. Chẳng bao lâu nữa, dữ liệu lớn còn cho biết liệu chúng ta có đang yêu đương gì không.

Khả năng của chúng ta trong việc làm những điều mới, làm nhiều hơn, tốt hơn, và nhanh hơn có thể để mở ra những giá trị

vô cùng to lớn, tạo ra người chiến thắng và những kẻ thất bại mới. Phần lớn giá trị của dữ liệu sẽ đến từ những ứng dụng phụ của nó, giá trị tương lai, chứ không chỉ đơn giản từ ứng dụng chính của nó, như chúng ta vẫn quen nghĩ. Kết quả là đối với hầu hết các loại dữ liệu, sẽ hợp lý khi thu thập nhiều nhất có thể và giữ lâu đến mức nào nó còn có thêm giá trị, và để cho những người khác phân tích nó nếu họ là người phù hợp hơn để tận dụng được giá trị của nó (miễn là có thể chia sẻ được các lợi ích mà việc phân tích mang lại).

Các công ty nào xác lập được vị trí của mình giữa những dòng chảy thông tin và thu thập được dữ liệu sẽ phát triển mạnh. Việc khai thác dữ liệu lớn một cách hiệu quả đòi hỏi phải có những kỹ năng kỹ thuật và rất nhiều trí tưởng tượng - một tư duy dữ-liệu-lớn. Nhưng cốt lõi của giá trị có thể về tay những người nắm giữ dữ liệu. Và đôi khi thông tin không chỉ là một tài sản quan trọng có thể nhìn thấy một cách rõ ràng, mà còn là dữ liệu xả được tạo ra bởi những tương tác của con người với thông tin. Một công ty thông minh có thể sử dụng chúng để cải thiện dịch vụ hiện có hoặc khởi động một dịch vụ hoàn toàn mới.



Phim minh họa của Oracle về cách khai thác dữ liệu lớn

Đồng thời, dữ liệu lớn mang đến cho chúng ta những rủi ro rất lớn. Nó vô hiệu hóa những cơ chế kỹ thuật và pháp lý cốt lõi mà thông qua đó chúng ta hiện đang cố gắng bảo vệ sự riêng tư. Trong quá khứ những gì cấu thành thông tin định danh cá nhân đều đã được biết - tên, số an sinh xã hội, hồ sơ thuế... - và do vậy tương đối dễ để bảo vệ. Ngày nay, ngay cả những dữ liệu vô hại

nhất cũng có thể tiết lộ nhân thân của ai đó nếu một nhà sưu tập dữ liệu đã tích lũy được đủ về nó. Việc ẩn danh hóa hoặc cách giấu tin thông thường không còn tác dụng. Hơn nữa, hiện nay việc nhắm mục tiêu vào một cá nhân để giám sát đòi hỏi một sự xâm phạm rộng lớn lên yếu tố riêng tư hơn bao giờ hết, bởi chính quyền không chỉ muốn biết nhiều thông tin nhất có thể về một người, mà còn muốn biết phạm vi rộng nhất về các mối quan hệ, các kết nối và tương tác.

Ngoài những thách thức về yếu tố riêng tư, những ứng dụng này của dữ liệu lớn còn làm nổi lên mối lo ngại đặc biệt và đáng ngại khác: nguy cơ chúng ta có thể đánh giá con người không chỉ với hành vi thực tế của họ mà với những khuynh hướng do dữ liệu cho thấy họ sẽ có. Khi các dự đoán dữ-liệu-lớn trở nên chính xác hơn, xã hội có thể sử dụng chúng để trừng phạt con người vì hành vi được dự đoán - những hành động mà họ chưa hề thực hiện. Những dự đoán như vậy là không thể bác bỏ một cách rõ ràng, vì vậy những người mà chúng cáo buộc không bao giờ có thể biện hộ được cho mình. Hình phạt trên cơ sở này phủ nhận nguyên lý tự do chí và bác bỏ khả năng, dù nhỏ bé tới đâu, rằng một con người có thể lựa chọn một con đường khác. Vì xã hội trao trách nhiệm cá nhân (và đưa ra hình phạt), ý chí của con người phải được xem là bất khả xâm phạm. Tương lai phải còn là một cái gì đó mà chúng ta có thể định hình theo thiết kế riêng của mình. Nếu không, dữ liệu lớn sẽ làm biến thái bản chất cốt lõi nhất của nhân loại: hợp lý trong suy nghĩ và tự do trong lựa chọn.

Không có cách rõ ràng nhất để chuẩn bị đầy đủ cho thế giới của dữ liệu lớn, nó sẽ đòi hỏi chúng ta thiết lập những nguyên tắc mới để chúng ta cai quản lấy chính mình. Một loạt thay đổi quan trọng đối với hoạt động của chúng ta có thể giúp ích cho xã hội khi nó trở nên quen thuộc hơn với đặc trưng và những thiếu sót của dữ liệu lớn. Chúng ta phải bảo vệ sự riêng tư bằng cách chuyển trách nhiệm khỏi các cá nhân và hướng tới những người

sử dụng dữ liệu - nghĩa là tới việc sử dụng có trách nhiệm. Trong một thế giới của các dự đoán, điều quan trọng là chúng ta phải đảm bảo ý chí con người được giữ bất khả xâm phạm, và chúng ta bảo vệ không chỉ quyền chọn lựa theo tiêu chuẩn đạo đức mà cả trách nhiệm cá nhân đối với những hành vi cá nhân.

Ngoài ra, xã hội phải thiết lập những biện pháp bảo vệ để giúp một chuyên ngành mới gồm các “nhà thuật toán” đánh giá các phân tích dữ-liệu-lớn - để một thế giới vốn trở nên ít ngẫu nhiên hơn do sự can thiệp của dữ liệu lớn không biến thành chiếc hộp đen. Nếu như vậy thì chẳng khác nào chuyển từ tình trạng mù mờ này sang tình trạng mù mờ khác.

Dữ liệu lớn sẽ được tích hợp vào quá trình tìm hiểu và giải quyết nhiều bài toàn toàn cầu mang tính cấp bách của chúng ta. Giải quyết vấn đề biến đổi khí hậu đòi hỏi phải phân tích dữ liệu ô nhiễm để biết nơi cần tập trung nỗ lực của chúng ta và tìm cách giảm thiểu các vấn nạn. Các cảm biến đang được đặt trên khắp thế giới, bao gồm cả những cái được nhúng trong các điện thoại thông minh, cung cấp vô số dữ liệu cho phép chúng ta mô hình hóa sự nóng lên toàn cầu ở mức độ chi tiết hơn. Trong khi đó, việc cải thiện và giảm chi phí chăm sóc sức khỏe, đặc biệt là cho người nghèo trên thế giới, phần lớn sẽ liên quan đến tự động hóa những công việc hiện tại dường như cần đánh giá của con người nhưng có thể được thực hiện bằng máy tính, chẳng hạn kiểm tra sinh thiết cho tế bào ung thư hoặc phát hiện nhiễm trùng trước khi các triệu chứng xuất hiện một cách đầy đủ.

Dữ liệu lớn đã được sử dụng cho sự phát triển kinh tế và ngăn ngừa xung đột. Nó đã phát hiện ra rằng những khu ổ chuột ở châu Phi là những cộng đồng sôi động về hoạt động kinh tế ,bằng cách phân tích các chuyển động của người sử dụng điện thoại di động. Nó đã phát hiện những khu vực chín muồi cho các cuộc đụng độ sắc tộc và chỉ ra các cuộc khủng hoảng tị nạn có thể xuất hiện như thế nào. Và các ứng dụng của nó chắc chắn nhân lên

khi công nghệ được áp dụng cho nhiều khía cạnh hơn của cuộc sống. Dữ liệu lớn giúp chúng ta làm tốt hơn những gì chúng ta đã làm, và cho phép chúng ta làm những điều hoàn toàn mới mẻ. Tuy nhiên, nó không phải là cây đũa thần. Nó sẽ không thể mang lại hòa bình thế giới, xóa bỏ đói nghèo, hoặc sản sinh một Picasso kế tiếp. Dữ liệu lớn không thể sinh ra một đứa bé - nhưng nó có thể cứu được những đứa trẻ bị sinh non. Rồi sẽ đến lúc chúng ta trông đợi nó được sử dụng trong hầu hết mọi khía cạnh của cuộc sống, và có lẽ chúng ta sẽ hoảng hốt một chút khi nó vắng mặt, giống như khi chúng ta mong một bác sĩ yêu cầu chụp X-quang để phát hiện các vấn đề có thể không phát hiện được khi khám bệnh.

Khi dữ liệu lớn trở nên phổ biến, nó cũng có thể ảnh hưởng đến cách chúng ta nghĩ về tương lai. Hiện tại có thể được định hình, còn tương lai đã chuyển từ một cái gì đó hoàn toàn dự đoán được thành một cái gì đó mở, nguyên sơ - một tấm vải bố rộng, trống trải mà mỗi cá nhân có thể vẽ lên theo những giá trị và nỗ lực của chính mình. Một trong những đặc điểm nổi bật của thời hiện đại là cảm giác tự chúng ta làm chủ số phận của mình - thái độ khiến chúng ta khác với tổ tiên. Tuy nhiên, dự đoán dữ-liệu-lớn khiến tương lai ít mở hơn và bị ảnh hưởng. Thay cho một tấm vải bạt trống, tương lai của chúng ta dường như đã được phác thảo bằng những dấu vết mờ nhạt có thể hiện lên rõ ràng bởi những người sở hữu công nghệ để làm rõ chúng. Điều này dường như làm giảm khả năng của chúng ta trong việc định hình số phận của mình. Trên bàn hành lễ của xác suất, năng lực tiềm ẩn của chúng ta chính là vật hiến tế.

Cùng lúc đó, dữ liệu lớn có thể khiến chúng ta mãi mãi là tù nhân của các hành động mà ta thực hiện trước kia. "Quá khứ là khúc dạo đầu", Shakespeare từng viết. Dù việc xấu hay việc tốt, dữ liệu lớn đều xét chúng trên cơ sở thuật toán. Liệu một thế giới của các dự đoán như thế có khiến chúng ta chán chường đến nỗi chẳng còn hứng thú chào đón bình minh, chẳng còn mong muốn đặt

dấu ấn nhân văn của mình trên thế giới?

Thật ra điều ngược lại sẽ khả thi hơn. Nếu đoán được các hành động diễn ra thế nào trong tương lai, chúng ta sẽ có thể thực hiện các bước khắc phục hậu quả để ngăn chặn các vấn đề hoặc cải thiện các kết quả. Chúng ta sẽ phát hiện những sinh viên đang bắt đầu trượt dốc sớm trước khi đến kỳ thi cuối cùng. Chúng ta sẽ phát hiện những ổ ung thư nhỏ xíu và điều trị chúng trước khi căn bệnh có cơ hội xuất hiện. Chúng ta sẽ thấy trước nguy cơ mang thai ở tuổi vị thành niên hoặc nguy cơ trở thành tội phạm và can thiệp để thay đổi kết quả được dự báo này, nhiều nhất trong khả năng của mình. Chúng ta sẽ ngăn chặn những vụ hỏa hoạn chết người trong những khu chung cư quá tải ở New York, nhờ biết những tòa nhà nào cần kiểm tra trước nhất.

Chẳng có gì được ấn định trước cả, bởi vì chúng ta luôn luôn có thể đáp ứng và phản ứng với những thông tin mình nhận được. Các dự đoán của dữ liệu lớn không phải được khắc ghi trên đá - chúng chỉ là những kết quả có khả năng xảy ra, và điều đó nghĩa là nếu muốn thay đổi, chúng ta có thể làm được. Chúng ta có thể xác định cách tốt nhất để chào đón tương lai và trở thành chủ nhân của nó, giống như Maury đã tìm thấy những tuyến đường tự nhiên trong không gian rộng mở của gió và sóng. Và để thực hiện điều này, chúng ta không bị buộc phải hiểu bản chất của vũ trụ hoặc chứng minh sự tồn tại của các vị thần - dữ liệu lớn là đủ tốt rồi.

Dữ liệu lớn hơn nữa

Khi dữ liệu lớn biến đổi cuộc sống của chúng ta - tối ưu hóa, cải thiện, tăng hiệu quả, và nắm bắt những lợi ích - vậy thì trực giác, đức tin, sự mơ hồ và tính độc đáo sẽ còn lại vai trò gì đây?

Nếu có điều gì dữ liệu lớn dạy cho chúng ta, đó chính là chỉ cần hành động tốt hơn, thực hiện những cải tiến, mà không cần hiểu

biết sâu sắc hơn; và thông thường như vậy là đủ rồi. Tiếp tục làm như vậy là đúng đắn. Thậm chí nếu bạn không biết tại sao những nỗ lực của mình lại hiệu quả, bạn vẫn đang tạo ra những kết quả tốt hơn so với khi bạn không tạo ra những nỗ lực như vậy. Flowers và "những đứa trẻ" của ông ở New York có thể không phải là hiện thân cho sự giác ngộ của các bậc thánh hiền, nhưng họ cứu được những mạng sống.

Dữ liệu lớn không phải là một thế giới lạnh lẽo của các thuật toán và máy tính. Vẫn có vai trò thiết yếu của con người, với tất cả những nhược điểm, nhận thức sai và lỗi lầm, bởi những đặc điểm đó đi song hành với sự sáng tạo, bản năng, và thiên tài của con người. Các quá trình hỗn độn tương tự trong tinh thần của chúng ta vốn dẫn đến định hướng sai, nhưng cũng dẫn đến những thành công và những ý tưởng vĩ đại thật tình cờ. Điều này cho thấy dù đang cố gắng nắm lấy thứ dữ liệu hỗn độn vì nó phục vụ một mục đích lớn hơn, chúng ta vẫn nên tiếp nhận sự không chính xác như một phần của nhân loại. Xét cho cùng, sự bừa bộn là một đặc tính cần thiết của cả thế giới và tâm thức chúng ta, và chúng ta chỉ được hưởng lợi bằng cách chấp nhận nó và áp dụng nó.

Cũng rất cần có một nơi cho con người để dành không gian cho trực giác, cho sự suy xét, nhằm đảm bảo chúng ta không bị dữ liệu và những câu trả lời bằng máy chôn vùi. Những điều tuyệt diệu nhất về con người chính là những điều các thuật toán và chip silicon không thể tiết lộ, bởi chúng không thể nắm bắt được trong dữ liệu.

Điều này có những hệ lụy quan trọng đối với quan niệm về sự tiến bộ trong xã hội. Dữ liệu lớn cho phép chúng ta thử nghiệm nhanh hơn và khám phá nhiều phương hướng hơn. Những lợi thế này đúng ra phải tạo nên nhiều bước đổi mới hơn. Tuy nhiên những tia sáng của phát minh lại là thứ mà dữ liệu không thể hiện được, dù với lượng dữ liệu lớn đến đâu chăng nữa, vì nó vẫn

chưa tồn tại. Nếu Henry Ford hỏi các thuật toán dữ-liệu-lớn rằng khách hàng của ông mong muốn gì, chúng sẽ trả lời: "một con ngựa nhanh hơn" (để nhắc lại câu nói nổi tiếng của ông). Trong một thế giới của dữ liệu lớn, chính những đặc điểm nhân văn nhất của chúng ta sẽ cần được khích lệ - sự sáng tạo, trực giác, và tham vọng tri thức - bởi vì tài khéo léo của chúng ta mới là nguồn gốc cho sự tiến bộ của nhân loại.

Dữ liệu lớn là một nguồn lực và một công cụ. Nó được tạo ra để thông báo, thay vì giải thích; nó dẫn chúng ta tới sự hiểu biết, nhưng nó vẫn có thể dẫn đến sự hiểu lầm, tùy thuộc vào việc nó được vận dụng tốt hay kém như thế nào. Và dù kinh ngạc đến đâu về sức mạnh của dữ liệu lớn, chúng ta không bao giờ được để sức quyến rũ của nó làm mình mù quáng đối với những khiếm khuyết vốn có của dữ liệu lớn.

Thông tin toàn vẹn về thế giới này - yếu tố tối thượng N = tất cả - sẽ chẳng bao giờ được thu thập, lưu trữ, hoặc xử lý bằng các công nghệ của chúng ta. Ví dụ phòng thí nghiệm vật lí hạt CERN ở Thụy Sĩ chỉ thu thập chưa đến 0,1 phần trăm các thông tin được tạo ra trong các thí nghiệm của nó - phần còn lại, dường như vô dụng, bị để bốc hơi vào hư vô. Nhưng điều đó khó có thể chấp nhận. Xã hội đã luôn luôn bị què quặt bởi những hạn chế của các công cụ chúng ta sử dụng nhằm đo lường và hiểu biết thực tế, từ la bàn, kính lục phân, rồi kính viễn vọng, radar tới GPS ngày nay. Các công cụ ngày mai của chúng ta có thể mạnh hơn gấp đôi, gấp mười hay gấp ngàn lần so với các công cụ của ngày hôm nay, khiến những gì chúng ta biết hôm nay có lẽ sẽ rất nhỏ khi đó. Thế giới dữ-liệu-lớn hiện tại của chúng ta chẳng bao lâu nữa sẽ trở nên kỳ quặc, cũng giống như 4 KB bộ nhớ cho phép ghi dữ liệu nằm trong máy tính điều khiển dẫn hướng Apollo 11 so với công nghệ của ngày hôm nay.

Những gì chúng ta có thể thu thập và xử lý sẽ luôn luôn chỉ là một phần nhỏ của các thông tin tồn tại trên thế giới. Nó chỉ có

thể là một hình ảnh của hiện thực, như những cái bóng trên tường trong cái hang của Plato. Bởi vì chúng ta không bao giờ có được thông tin hoàn hảo, nên các dự đoán của chúng ta vốn dĩ luôn có thể sai lầm. Điều này không có nghĩa chúng là sai, chỉ là chúng luôn luôn không đầy đủ. Nó không phủ nhận những hiểu biết mà dữ liệu lớn cung cấp, nhưng nó đặt dữ liệu lớn vào đúng vị trí của nó - một công cụ không cung cấp các câu trả lời cuối cùng, mà chỉ những câu trả lời *đủ tốt* để giúp chúng ta bây giờ cho đến khi có được các phương pháp tốt hơn, và cùng với đó là các câu trả lời tốt hơn. Nó cũng cho thấy rằng chúng ta phải sử dụng công cụ này với rất nhiều sự khiêm nhường... và cả tính nhân văn nữa.

CHÚ GIẢI THÔNG TIN

1. HIỆN TẠI

Xu hướng Dịch cúm của Google - Jeremy Ginsburg et al., “Detecting Influenza Epidemics Using Search Engine Query Data”, *Nature* 457 (2009), pp. 1012-14 (http://www.nature.com/ nature/journal/v457/n7232/full/nature07634.html). Nghiên cứu tiếp theo về Xu hướng Dịch cúm của Google - A. F. Dugas et al., “Google Flu Trends: Correlation with Emergency Department Influenza Rates and Crowding Metrics”, CID Advanced Access (January 8, 2012); DOI 10.1093 /cid/cir883. Mua vé máy bay, Farecast - Các thông tin xuất phát từ Kenneth Cukier, “Data, Data Everywhere”, *The Economist* special report, February 27, 2010, pp. 1-14, và từ các cuộc phỏng vấn với Etzioni giữa năm 2010 và 2012.

Dự án Hamlet của Etzioni - Oren Etzioni, C.A. Knoblock, R. Tuchinda, and A. Yates, “To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price”, SIGKDD ’03, August 24-27, 2003 (http://knight.cis.temple.edu/~yates//papers/ hamlet-kdd03.pdf).

Giá Microsoft trả cho Farecast - Từ báo cáo truyền thông, đặc biệt là “Secret Farecast Buyer Is Microsoft”, Seattlepi.com, April 17, 2008 (http://blog.seattlepi.com/venture/2008/04/17/secretfarecast-buyer-is-microsoft/?source=mypi).

Một cách nghĩ về dữ liệu lớn - Có một cuộc tranh luận ồn ào nhưng không hiệu quả về nguồn gốc của thuật ngữ “dữ liệu lớn” và làm sao để định nghĩa nó một cách hoàn hảo. Thuật ngữ này đã thỉnh thoảng xuất hiện từ nhiều thập kỷ nay. Một báo cáo

nghiên cứu năm 2001 bởi Doug Laney của Gartner đưa ra công thức “ba V” của dữ liệu lớn (volume, velocity, variety - khối lượng, vận tốc, và tính đa dạng), tuy hữu ích vào lúc đó nhưng không hoàn hảo. Thiên văn học và xác định trình tự DNA - Cukier, “Data, Data Everywhere”. Hàng tỷ cổ phiếu được mua bán - R ita Nazareth and Julia Leite, “Stock Trading in U.S. Falls to Lowest Level Since 2008”, *Bloomberg,* August 13, 2012 (http://www.bloomberg.com/news/2012-08-13/stock-trading-in-u-s-hits-lowest-levelsince-2008-as-vix-falls.html).

24 petabyte mỗi ngày của Google - Thomas H. Davenport, Paul Barth, and Randy Bean, “How ‘Big Data’ Is Different”, *Sloan Review,* July 30, 2012, pp. 43-46 (http://sloanreview.mit.edu/ the- magazine/ 2012- fall/ 54104/how- big- data- is- different/). Số liệu thống kê Facebook - Facebook IPO prospectus, “Form S-1 Registration Statement”, U.S. Securities and Exchange Commission, February 1, 2012 (http://sec.gov/Archives/edgar/ data/1326801/000119312512034517/d287954ds1.htm). Số liệu thống kê YouTube - Larry Page, “Update from the CEO”, Google, April 2012 (http://investor.google.com/corporate/2012/ ceo-letter.html).

Số lượng tweet - Tomio Geron, “Twitter’s Dick Costolo: Twitter Mobile Ad Revenue Beats Desktop on Some Days”, *Forbes,* June 6, 2012 (http://www.forbes.com/sites/tomiogeron/2012/06/06/ twitters-dick-costolo-mobile-ad-revenue-beats-desktop-onsome-days/).

Thông tin về số lượng của dữ liệu - Martin Hilbert and Priscilla López, “The World’s Technological Capacity to Store, Communicate, and Compute Information” *Science,* April 1, 2011, pp. 60-65; Martin Hilbert and Priscilla López, “How to Measure the World’s Technological Capacity to Communicate, Store and Compute Information?” *International Journal of Communication* 2012, pp. 1042-55 (http://www.ijoc.org/ojs/

index.php/ijoc/article/viewFile/1562/742).

Ước tính lượng thông tin được lưu trữ vào năm 2013 - Cukier phỏng vấn Hilbert, 2012.

In ấn và tám triệu cuốn sách; xuất bản nhiều hơn kể từ khi thành lập Constantinople - E lizabeth L. Eisenstein, *The Printing Revolution in Early Modern Europe* (Canto/Cambridge University Press, 1993), pp. 13-14.

Phép tương tự của Peter Norvig - Từ các buổi nói chuyện của Norvig dựa vào bài: A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data", *IEEE Intelligent Systems,* March/April 2009, pp. 8-12 (http://www.computer.org/portal/cms_docs_intelligent/intelligent/homepage/2009/x2exp.pdf). (Lưu ý rằng tiêu đề là từ bài viết của Eugene Wigner "Tính Hiệu quả phi lý của Toán học trong Khoa học Tự nhiên", trong đó ông xem xét tại sao vật lý có thể được thể hiện rất đẹp trong toán học cơ bản nhưng khoa học xã hội lại chống lại những công thức gọn gàng như vậy. Xem E. Wigner "The Unreasonable Effectiveness of Mathematics in the Natural Sciences," *Communications on Pure and Applied Mathematics* 13, no. 1 (1960), pp. 1-14.) "Peter Norvig - The Unreasonable Effectiveness of Data", lecture at University of British Columbia, YouTube, September 23, 2010 (http://www.youtube.com/watch?v=yvDCzhbjYWs). Về kích thước vật lý ảnh hưởng đến định luật vật lý thực hành (mặc dù không hoàn toàn chính xác), nguồn tham khảo thường được trích dẫn là J. B. S. Haldane, "On Being the Right Size", *Harper's Magazine,* March 1926 (http://harpers.org/archive/1926/03/onbeing-the-right-size/).

Picasso và những hình ảnh Lascaux - David Whitehouse, "UK Science Shows Cave Art Developed Early", *BBC News Online,* October 3, 2001 (http://news.bbc.co.uk/1/hi/sci/tech/1577421.stm).

2. NHIỀU HƠN

Trích dẫn Jeff Jonas - Conversation with Jonas, December 2010, Paris.

Lịch sử của điều tra dân số Mỹ - U.S. Census Bureau, “The Hollerith Machine” Online history. (http://www.census.gov/history/www/innovations/technology/the_hollerith_tabulator.html.

Đóng góp của Neyman - William Kruskal and Frederick Mosteller, “Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939”, *International Statistical Review* 48 (1980), pp. 169-195, pp. 187-188. Bài viết nổi tiếng của Neyman là “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection”, *Journal of the Royal Statistical Society* 97, no. 4 (1934), pp. 558-625.

Một mẫu của 1.100 quan sát là đủ - Earl Babbie, *Practice of Social Research* (12th ed. 2010), pp. 204-207.

Tác dụng của điện thoại di động - “Estimating the Cellphone Effect”, September 20, 2008 (http://www.fivethirtyeight.com/2008/09/estimating-cellphone-effect-22-points.html); để biết thêm về những định kiến trong việc bỏ phiếu và những hiểu biết thống kê khác, xem Nate Silver, *The Signal and the Noise: Why So Many Predictions*

Trình tự gen của Steve Jobs - Walter Isaacson, *Steve Jobs* (Simon and Schuster, 2011), pp. 550-551.

Xu hướng Dịch cúm Google dự đoán đến cấp thành phố - Dugas et al., “Google Flu Trends”.

Etzioni về dữ liệu thời gian - Interview by Cukier, October 2011. Trích dẫn John Kunze - Jonathan Rosenthal, “Special Report: International Banking”, *The Economist,* May 19, 2012, pp. 7-8.

Gian lận các trận đấu sumo - Mark Duggan and Steven D. Levitt, “Winning Isn’t Everything: Corruption in Sumo Wrestling”, *American Economic Review* 92 (2002), pp. 1594-1605 (http://pricetheory.uchicago.edu/levitt/Papers/DugganLevitt2002.pdf).

11 triệu tia ánh sáng của Lytro - từ trang web của công ty Lytro (http://www.lytro.com).

Thay thế lấy mẫu trong khoa học xã hội - Mike Savage and Roger Burrows, “The Coming Crisis of Empirical Sociology”, *Sociology* 41 (2007), pp. 885-899.

Về phân tích dữ liệu toàn diện từ một nhà điều hành điện thoại di động - J. P. Onnela et al., “Structure and Tie Strengths in Mobile Communication Networks”, *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 104 (May 2007), pp. 7332-36 (http://nd.edu/~dddas/Papers/PNAS0610245104v1.pdf).

3. HỖN ĐỘN

Crosby - Alfred W. Crosby, The Measure of Reality: Quantification and Western Society, 1250-1600 *(Cambridge University Press, 1997).*

Về các trích dẫn của Kelvin và Bacon - Những câu cách ngôn này được nhiều người cho là của hai ông, mặc dù phát biểu thực tế trong tác phẩm viết của họ hơi khác. Với Kelvin, nó là một phần của một trích dẫn về đo lường, từ bài giảng của ông tên là “Electrical Units of Measurement” (1883). Với Bacon, nó được

xem là một bản dịch chưa chặt chẽ từ tiếng Latin, trong *Meditationes Sacrae* (1597).

Nhiều cách để hiểu từ viết tắt IBM - DJ Patil, “Data Jujitsu: The Art of Turning Data into Product”, *O’Reilly Media,* July 2012 (http://oreillynet.com/oreilly/data/radarreports/data-jujitsu.csp?cmp=tw-strata-books-data-products).

30.000 giao dịch mỗi giây trên NYSE - Colin Clark, “Improving Speed and Transparency of Market Data”, NYSE EURONEX T blog post, January 9, 2011 (http://exchanges.nyx.com/cclark/improving-speed-and-transparency-market-data).

Ý tưởng “2 + 2 = 3,9” - Brian Hopkins and Boris Evelson, “Expand Your Digital Horizon with Big Data”, Forrester, September 30, 2011.

Những cải thiện trong các thuật toán - President’s Council of Advisors on Science and Technology, “Report to the President and Congress, Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology”, December 2010, p. 71 (http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrdreport-010.pdf).

Các thế cờ tàn - Bảng thế cờ tàn toàn diện nhất được công bố, bảng Nalimov (đặt theo tên của một trong những người lập ra nó), bao gồm tất cả các ván cờ cho sáu quân cờ hoặc ít hơn. Dung lượng của nó là hơn 7 terabyte, và việc nén thông tin trong đó là một thách thức lớn. Xem E. V. Nalimov, G. McC. Haworth, and E. A. Heinz, “Space-efficient Indexing of Chess Endgame Tables”, *ICGA Journal* 23, no. 3 (2000), pp. 148-162.

Microsoft và hiệu suất thuật toán - Michele Banko and Eric Brill, “Scaling to Very Very Large Corpora for Natural Language Disambiguation”, Microsoft Research, 2001, p. 3 (http://acl.ldc.

upenn.edu/P/P01/P01-1005.pdf).

Bản thử nghiệm, lời nói, và trích dẫn của IBM - IBM, “701 Translator”, press release, IBM archives, January 8, 1954 (http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html). Xem thêm John Hutchins, “The First Public Demonstration of Machine Translation: The Georgetown-IBM System, 7th January 1954”, November 2005 (http://www.hutchinsweb. me.uk/GU-IBM-2005.pdf).

IBM Candide - Adam L. Berger et al., “The Candide System for Machine Translation”, *Proceedings of the 1994 ARPA Workshop on Human Language Technology,* 1994 (http://aclweb.org/anthology-new/H/H94/H94-1100.pdf).

Lịch sử của dịch thuật máy - Yorick Wilks, *Machine Translation: Its Scope and Limits* (Springer, 2008), p. 107.

Hàng triệu văn bản của Candide so với hàng tỷ văn bản của Google - Och interview with Cukier, December 2009.

Tập sao lục 95 tỷ câu của Google - Alex Franz and Thorsten Brants, “All Our N-gram are Belong to You”, Google blog post, August 3, 2006 (http://googleresearch.blogspot.co.uk/2006/08/all-our-n-gram-are-belong-to-you.html).

Tập sao lục Brown và 1 nghìn tỷ từ của Google - Halevy, Norvig, and Pereira, “The Unreasonable Effectiveness of Data”.

Trích dẫn từ bài viết của đồng tác giả Norvig - sđd.

Sự ăn mòn đường ống của BP và môi trường không dây gây hại - Jaclyn Clarabut, “Operations Making Sense of Corrosion”, *BP Magazine,* issue 2 (2011) (http://www.bp.com/liveassets/ bp_internet/globalbp/globalbp_uk_english/reports_and_publications/bp_magazine/STAGING/local_assets/pdf/BP_Maga

Khó khăn trong việc đọc dữ liệu không dây - Cukier, “Data, Data, Everywhere”. Hệ thống này rõ ràng không thể sai lầm: một đám cháy tại nhà máy lọc dầu BP Cherry Point vào tháng 2 năm 2012 được quy lỗi cho một đường ống bị ăn mòn.

Dự án với giá hàng tỷ - Từ cuộc phỏng vấn với người đồng sáng lập với Cukier, Tháng 10 năm 2012. James Surowiecki, “A Billion Prices Now”, *The New Yorker,* May 30, 2011; dữ liệu và các chi tiết có thể được tìm thấy trên trang web của dự án (http://bpp. mit.edu/); Annie Lowrey, “Economists’ Programs Are Beating U.S. at Tracking Inflation”, *Washington Post,* December 25, 2010 (http://www.washingtonpost.com/wp-dyn/content/ article/2010/12/25/AR2010122502600.html).

Price Stats với vai trò kiểm tra số liệu thống kê quốc gia - “Official Statistics: Don’t Lie to Me, Argentina”, *The Economist,* February 25, 2012 (http://www.economist.com/node/21548242). Số lượng hình ảnh trên Flickr - Từ trang web Flickr (http://www. flick.com).

Về thách thức đối với phân loại thông tin - David Weinberger, *Everything Is Miscellaneous: The Power of the New Digital Disorder* (Times, 2007).

Pat Helland - Pat Helland, “If You Have Too Much Data Then ‘Good Enough’ Is Good Enough”, *Communications of the ACM,* June 2011, pp. 40, 41. Có một cuộc tranh luận sôi nổi trong cộng đồng cơ sở dữ liệu về các mô hình và khái niệm tốt nhất có thể để đáp ứng các nhu cầu của dữ liệu lớn. Helland đại diện cho nhóm để nghị bỏ các công cụ đã được sử dụng trong quá khứ. Michael Rys, “Scalable SQL”, *Communications of the ACM,* June 2011, p. 48. Bài này cho rằng những phiên bản được áp dụng nhiều của các công cụ hiện có sẽ làm việc tốt.

Visa sử dụng Hadoop - Cukier, “Data, data everywhere”. Chỉ có 5 phần trăm thông tin là dữ liệu có cấu trúc - Abhishek Mehta, “Big

Data: Powering the Next Industrial Revolution", Tableau Software White Paper, 2011 (http://www.tableausoftware.com/learn/whitepapers/big-data-revolution).

4. TƯƠNG QUAN

Câu chuyện của Linden cũng như "tiếng nói của Amazon" - Linden interview with Cukier, March 2012.

WSJ trong các bài phê bình trên Amazon - Như trích dẫn trong James Marcus, *Amazonia: Five Years at the Epicenter of the Dot. Com Juggernaut* (New Press, 2004), p. 128.

Trích dẫn Marcus - Marcus, Amazonia, p. 199.

Các giới thiệu là một phần ba thu nhập của Amazon - Con số này chưa bao giờ được công ty chính thức xác nhận nhưng đã được xuất bản trong nhiều báo cáo phân tích và bài viết trên phương tiện truyền thông, bao gồm cả "Building with Big Data: The Data Revolution Is Changing the Landscape of Business", *The Economist,* May 26, 2011 (http://www.economist.com/node/18741392/).

Con số này cũng đã được tham chiếu bởi hai cựu giám đốc điều hành Amazon trong các cuộc phỏng vấn với Cukier.

Thông tin giá Netflix - Xavier Amatriain and Justin Basilico, "Netflix Recommendations: Beyond the 5 stars (Part 1)", Netflix blog, April 6, 2012.

"Bị lừa bởi Ngẫu nhiên" - Nassim Nicholas Taleb, *Fooled by Randomness* (Random House, 2008); Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable* (2nd ed., Random House, 2010).

Walmart và Pop-Tarts - Constance L. Hays, "What Wal-Mart

Knows About Customers' Habits", *New York Times,* November 14, 2004 (http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html).

Ví dụ về các mô hình dự báo của FICO, Experian, và Equifax - Scott Thurm, "Next Frontier in Credit Scores: Predicting Personal Behavior", *Wall Street Journal,* October 27, 2011 (http://online.wsj.com/article/SB10001424052970203687504576655182086912.html).

Các mô hình dự báo của Aviva - Leslie Scism and Mark Maremont, "Insurers Test Data Profiles to Identify Risky Clients", *Wall Street Journal,* November 19, 2010 (http://online.wsj.com/article/SB10001424052748704648604575620750998986.html); Leslie Scism and Mark Maremont, "Inside Deloitte's Life-Insurance Assessment Technology", *Wall Street Journal,* November 19, 2010 (http://online.wsj.com/article/SB10001424052748704104104575622531084755588.html); Howard Mills, "Analytics: Turning Data into Dollars", *Forward Focus,* December 2011 (http://www.deloitte.com/assets/Dcom-UnitedStates/
Local%20Assets/Documents/FSI/US_FSI_Forward%20Focus_
Analytics_ Turning%20data%20into%20dollars_120711.pdf).
Ví dụ về Target và thiếu niên mang thai - Charles Duhigg, "How Companies Learn Your Secrets", *New York Times,* February 16, 2012 (http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html). Bài viết được chuyển thể từ cuốn sách của Duhigg, *The Power of Habit: Why We Do What We Do in Life and Business* (Random House, 2012); Target đã tuyên bố có những sự thiếu chính xác trong báo cáo của phương tiện truyền thông về các hoạt động của mình nhưng từ chối cho biết chúng là những gì. Khi được hỏi về vấn đề với cuốn sách này, một phát ngôn viên của Target trả lời: "Mục đích là sử dụng dữ liệu khách hàng để tăng cường mối quan hệ của khách hàng với Target. Khách hàng của chúng tôi muốn nhận được giá trị cao, những lời

chào hàng thích hợp, và một trải nghiệm vượt trội. Giống như nhiều công ty, chúng tôi sử dụng công cụ nghiên cứu giúp hiểu được xu hướng mua sắm và sở thích của khách hàng để có thể gửi lời chào hàng và chương trình khuyến mãi phù hợp với họ. Chúng tôi có trách nhiệm bảo vệ lòng tin của khách hàng một cách rất nghiêm túc. Một trong những cách chúng tôi áp dụng là có một chính sách bảo mật toàn diện mà chúng tôi chia sẻ công khai trên Target.com, và thường xuyên dạy các nhân viên của chúng tôi cách bảo vệ thông tin của khách hàng".

Các phân tích của UPS tỏ ra hiệu quả - Cukier interviews with Jack Levis, 2012.

Trẻ sinh thiếu tháng - Dựa trên các cuộc phỏng vấn với McGregor trong năm 2010và năm 2012. Carolyn McGregor, Christina Catley, Andrew James, và James Padbury, "Next Generation Neonatal Health Informatics with Artemis", in European Federation for Medical Informatics, *User Centred Networked Health Care,* ed. A. Moen et al. (IOS Press, 2011), p. 117. Một số tài liệu xuất phát từ Cukier, "Data, Data, Everywhere".

Về tương quan giữa hạnh phúc và thu nhập - R. Inglehart and H.-D. Klingemann, *Genes, Culture and Happiness* (MIT Press, 2000).

Về bệnh sởi và các chi phí y tế, cùng các công cụ phi tuyến tính mới cho phân tích tương quan - David Reshef et al., "Detecting Novel Associations in Large Data Sets", *Science* 334 (2011), pp. 1518-24.

Kahneman - Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011), pp. 74-75.

Pasteur - Đối với độc giả quan tâm đến ảnh hưởng lớn của Pasteur về cách chúng ta nhận thức sự vật, mời xem Bruno Latour, *The Pasteurization of France* (Harvard University Press, 1993). Nguy cơ mắc bệnh dại - Melanie Di Quinzio and Anne

McCarthy, “Rabies Risk Among Travellers”, *CMAJ* 178, no. 5 (2008), p. 567. Nhân quả hiếm khi có thể được chứng minh - Nhà khoa học máy tính đoạt giải thưởng Turing, Judea Pearl, đã phát triển một cách để chính thức thể hiện động lực quan hệ nhân quả; dù không có bằng chứng chính thức, điều này cung cấp một cách tiếp cận thực tế để phân tích các quan hệ nhân quả. Judea Pearl, *Causality: Models, Reasoning and Inference* (Cambridge University Press, 2009).

Ví dụ xe Orange - Quentin Hardy. “Bizarre Insights from Big Data”, nytimes.com, March 28, 2012 (http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/); and Kaggle, “Momchil Georgiev Shares His Chromatic Insight from Don’t Get Kicked”, blog posting, February 2, 2012 (http://blog.kaggle.com/2012/02/02/momchil-georgiev-shares-his-chromaticinsight-from-dont-get-kicked/).

Sức nặng của nắp cống, số lượng các vụ nổ, và chiều cao của các vụ nổ - R achel Ehrenberg, “Predicting the Next Deadly Manhole Explosion”, *Wired,* July 7, 2010 (http://www.wired.com/wiredscience/2010/07/manhole-explosions).

Con Edison làm việc với các nhà thống kê thuộc Đại học Columbia - trường hợp này được mô tả cho độc giả trong Cynthia Rudin et al., “21st-Century Data Miners Meet 19th-Century Electrical Cables”, *Computer,* June 2011, pp. 103-105. Các mô tả kỹ thuật của công trình có trong những bài báo chuyên ngành của Rudin và cộng sự trên các trang web của họ, đặc biệt là Cynthia Rudin et al., “Machine Learning for the New York City Power Grid”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, no. 2 (2012), pp. 328-345 (http://hdl.handle.net/1721.1/68634).

Sự hỗn độn của thuật ngữ “tủ điện” - Rudin et al., “21st-Century Data Miners Meet 19th-Century Electrical Cables”.

Trích dẫn của Rudin từ cuộc phỏng vấn với Cukier, tháng 3 năm 2012.

Các lượt xem của Anderson - Chris Anderson, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”, *Wired,* June 2008 (http://www.wired.com/science/discoveries/magazine/16-07/pb_theory/).

Anderson rút lại tuyên bố - National Public Radio, “Search and Destroy”, July 18, 2008 (http://www.onthemedia.org/2008/jul/18/search-and-destroy/transcript/).

Về các lựa chọn ảnh hưởng đến phân tích của chúng ta - danah boyd and Kate Crawford. “Six Provocations for Big Data”, paper presented at Oxford Internet Institute’s “A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society”, September 21, 2011 (http://ssrn.com/abstract=1926431).

5. DỮ LIỆU HÓA

Các chi tiết về cuộc sống của Maury được biên soạn từ nhiều tác phẩm của ông và về ông: Chester G. Hearn, *Tracks in the Sea: Matthew Fontaine Maury and the Mapping of the Oceans* (International Marine/McGraw-Hill, 2002); Janice Beaty, *Seeker of Seaways: A Life of Matthew Fontaine Maury, Pioneer Oceanographer* (Pantheon Books, 1966); Charles Lee Lewis, *Matthew Fontaine Maury: The Pathfinder of the Seas* (U.S. Naval Institute, 1927) (http://archive.org/details/matthewfontainem00lewi); Matthew Fontaine Maury, *The Physical Geography of the Sea* (Harper, 1855).

Trích dẫn của Maury - Maury, *Physical Geography of the Sea,* “Introduction”, pp. xii, vi.

Dữ liệu về ghế xe hơi - Nikkei, “Car Seat of Near Future IDs Driver’s Backside”, December 14, 2011.

Đinh lượng thế giới - Phần lớn suy nghĩ của tác giả về lịch sử dữ liệu hóa đã được lấy cảm hứng từ Crosby, *The Measure of Reality.* Người châu Âu chưa bao giờ được tiếp xúc với bàn tính - Sđd, 112. Calculating faster using Arabic numerals – Alexander Murray, *Reason and Society in the Middle Ages* (Oxford University Press, 1978), p. 166.

Tổng số sách được xuất bản và nghiên cứu của Harvard về dự án sao chụp sách của Google - Jean-Baptiste Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books", *Science* 331 (January 14, 2011), pp. 176-182 (http://www.
sciencemag.org/content/331/6014/176.abstract). Về bài giảng video - Erez Lieberman Aiden and Jean-Baptiste Michel, "What We Learned from 5 Million Books", TEDx, Cambridge, MA, 2011 (http://www.ted.com/talks/what_we_learned_from_5_million_books.html).

Về các mô-đun vô tuyến trong xe hơi và bảo hiểm - Cukier, "Data, Data Everywhere".

Jack Levis của UPS - Interview with Cukier, April 2012.

Số liệu về khoản tiết kiệm được của UPS - Institute for Operations Research and the Management Sciences (INFORMS), "UPS Wins Gartner BI Excellence Award", 2011 (http://www.informs.org/ Announcements/UPS-wins-Gartner-BI-Excellence-Award). Nghiên cứu Pentland - Robert Lee Hotz, "The Really Smart Phone", *Wall Street Journal,* April 22, 2011 (http://online.wsj.
com/article/SB10001424052748704547604576263261679848 4.html).

Nghiên cứu các khu ổ chuột của Eagle - Nathan Eagle, "Big Data, Global Development, and Complex Systems", Santa Fe Institute,

May 5, 2010 (http://www.youtube.com/watch?v=yaivtqlu7iM); Interview with Cukier, October 2012.

Dữ liệu Facebook - Facebook IPO Prospectus, 2012.

Dữ liệu Twitter - Alexia Tsotsis, "Twitter Is at 250 Million Tweets per Day, iOS 5 Integration Made Signups Increase 3x", *TechCrunch,* October 17, 2011, http://techcrunch.com/2011/10/17/twitter-is-at-250-million-tweets-per-day/.
Quỹ phòng hộ sử dụng Twitter - Kenneth Cukier, "Tracking Social Media: The Mood of the Market", *Economist.com,* June 28, 2012
(http://www.economist.com/blogs/graphicdetail/2012/06/tracking-social-media).

Twitter và dự báo doanh thu phòng vé của Hollywood - Sitaram Asur and Bernardo A. Huberman, "Predicting the Future with Social Media", *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,* pp. 492-499; online at http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf.
Twitter và cảm xúc trên toàn cầu - Scott A. Golder and Michael W. Macy, "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures", *Science* 333 (September 30, 2011), pp. 1878-81.

Twitter và tiêm phòng cúm - Marcel Salathé and Shashank Khandelwal, "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control", *PLoS Computational Biology,* October 2011.

Bằng sáng chế "Sàn thông minh" của IBM - Lydia Mai Do, Travis M. Grigsby, Pamela Ann Nesbitt, and Lisa Anne Seacat. "Securing premises using surfaced-based computing technology", U.S. Patent number: 8138882. Issue date: March 20, 2012. Phong trào tự định lượng - "Counting Every Moment", *The Economist,*

March 3, 2012.

Tai nghe của Apple cho các phép đo sinh học - Jesse Lee Dorogusker, Anthony Fadell, Donald J. Novotney, and Nicholas R Kalayjian, "Integrated Sensors for Tracking Performance Metrics", U.S. Patent Application 20090287067. Assignee: Apple. Application Date: 2009-07-23. Publication Date: 2009-11-19. Sinh trắc học Derawi - "Your Walk Is Your PIN-Code", press release, February 21, 2011 (http://biometrics.derawi. com/?p=175).

Thông tin iTrem - The Landmarc Research Center at Georgia Tech (http://eosl.gtri.gatech.edu/Capabilities/LandmarcResearchCent tabid/798/Default.aspx).

Các nhà nghiên cứu Kyoto về gia tốc ba trục - iMedicalApps Team, "Gait Analysis Accuracy: Android App Comparable to Standard Accelerometer Methodology", *mHealth,* March 23, 2012. Báo chí đã thúc đẩy nhà nước độc lập - Benedict Anderson, *Imagined Communities: Reflections on the Origin and Spread of Nationalism* (Verso, 2006).

Các nhà vật lý cho thấy thông tin là cơ sở của tất cả mọi thứ - Hans Christian von Baeyer, *Information: The New Language of Science* (Harvard University Press, 2005).

6. GIÁ TRỊ

Câu chuyện của Luis von Ahn được dựa trên các cuộc phỏng vấn của Cukier với von Ahn từ năm 2010. Xem thêm Clive Thompson, "For Certain Tasks, the Cortex Still Beats the CPU", *Wired,* June 25, 2007 (http://www.wired.com/techbiz/it/magazine/15-07/ff_humancomp?currentPage=all); Jessie Scanlon, "Luis von Ahn: The Pioneer of 'Human Computation,' "

Businessweek, November 3, 2008 (http://www.businessweek.com/stories/2008-11-03/luis-von-ahn-the-pioneer-of-humancomputation-businessweek-business-news-stock-market-andfinancial-advice). Mô tả kỹ thuật về reCaptchas - Luis von Ahn et al., “reCAPTCHA: Human-Based Character Recognition via Web Security Measures”, *Science* 321 (September 12, 2008), pp. 1465-68 (http://www.sciencemag.org/content/321/5895/1465.abstract).

Nhà máy sản xuất pin của Smith - Adam Smith, *The Wealth of Nations* (reprint, Bantam Classics, 2003), book I, chapter one. (Phiên bản điện tử miễn phí tại http://www2.hn.psu.edu/faculty/jmanis/adam-smith/Wealth-Nations.pdf).

Lưu trữ - Viktor Mayer-Schönberger, *Delete: The Virtue of Forgetting in the Digital Age* (Princeton University Press, 2011), p. 63.

Về sử dụng năng lượng của xe hơi điện - IBM, “IBM, Honda, and PG&E Enable Smarter Charging for Electric Vehicles”, press release, April 12, 2012 (http://www-03.ibm.com/press/us/en/pressrelease/37398.wss). Xem thêm Clay Luthy, “Guest Perspective: IBM Working with PG&E to Maximize the EV Potential” *PGE Currents* Magazine, April 13, 2012 (http://www.pgecurrents.com/2012/04/13/ibm-working-with-pge-tomaximize-the-ev-potential).

Amazon và dữ liệu của AOL - Cukier interviews with Andreas Weigend, 2010 and 2012.

Phần mềm Nuance và Google - Cukier, “Data, Data Everywhere”. Công ty Logistics - Brad Brown, Michael Chui, and James Manyika, “Are You Ready for the Era of ‘Big Data’?” *McKinsey Quarterly,* October 2011, p. 10.

Telefonica kiếm tiền với thông tin điện thoại di động - "Telefonica Hopes 'Big Data' Arm Will Revive Fortunes", *BBC Online,* October 9, 2012. (http://www.bbc.co.uk/news/technology-19882647).

Nghiên cứu của Hiệp hội Ung thư Đan Mạch - Patrizia Frei et al., "Use of Mobile Phones and Risk of Brain Tumours: Update of Danish Cohort Study", *BMJ* 343 (2011) (http://www.bmj.com/content/343/bmj.d6387), and interview with Cukier, October 2012. Dữ liệu GPS và xe tự hành Street View của Google - Peter Kirwan, "This Car Drives Itself", *Wired UK,* January 2012 (http://www.wired.co.uk/magazine/archive/2012/01/features/thiscar-drives-itself?page=all).

Về chương trình kiểm tra chính tả của Google và các trích dẫn - Interview with Cukier at the Googleplex in Mountain View, California, December 2009; Cukier, "Data, Data Everywhere". Sự sáng suốt của Hammerbacher - Interview with Cukier, October 2012.

Dữ liệu e-book của Barnes & Noble - Alexandra Alter, "Your E-Book Is Reading You", *Wall Street Journal,* June 29, 2012 (http://online.wsj.com/article/SB1000142405270230487030457490950051438304.html).

Lớp học và dữ liệu Coursera của Andrew Ng - Interview with Cukier, June 2012.

Chính sách chính phủ mở của Obama - Barack Obama, "Presidential memorandum", White House, January 21, 2009. Về giá trị dữ liệu của Facebook - Doug Laney, "To Facebook You're Worth $80.95", *Wall Street Journal,* May 3, 2012 (http://blogs.wsj.com/cio/2012/05/03/to-facebook-youreworth-80-95/).

Để định giá các mục tin rời rạc của Facebook, Laney ngoại suy từ

tốc độ tăng trưởng của Facebook để ước tính 2,1 nghìn tỷ mẩu nội dung. Trong bài viết trên WSJ của mình, ông định giá mỗi mục tin là 3 cent vì ông sử dụng ước tính giá trị thị trường trước đó của Facebook là 75 tỷ USD. Cuối cùng, nó là hơn 100 tỷ USD, hay 5 cent, như chúng ta ngoại suy dựa trên tính toán của ông. Khoảng cách giá trị của tài sản hữu hình và vô hình - Steve M. Samek, "Prepared Testimony: Hearing on Adapting a 1930's Financial Reporting Model to the 21st Century", U.S. Senate Committee on Banking, Housing and Urban Affairs, Subcommittee on Securities, July 19, 2000.

Giá trị của tài sản vô hình - Robert S. Kaplan and David P. Norton, *Strategy Maps: Converting Intangible Assets into Tangible Outcomes* (Harvard Business Review Press, 2004), pp. 4-5. Trích dẫn của Tim O'Reilly - Interview with Cukier, February 2011.

7. NHỮNG TÁC ĐỘNG

Thông tin về Decide.com được lấy từ các trao đổi email của Cukier với Etzioni vào tháng 5 năm 2012.

Báo cáo McKinsey - James Manyika et al., "Big Data: The Next Frontier forInnovation, Competition, and Productivity", McKinsey Global Institute,May 2011 (http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation), p. 10.

Trich dẫn của Hal Varian - Interview with Cukier, December 2009.

Trich dẫn của Carl de Marcken được lấy từ các trao đổi email với Cukier vào tháng 5 năm 2012.

Về MasterCard Advisors - Cukier interviews with Gary Kearns, *The Economist's* "The Ideas Economy: Information" conference,

Santa Clara, California, June 8, 2011.

Thông tin về Accenture và thành phố St Louis, Missouri được lấy từ bài phỏng vấn của Cukier với nhân viên thành phố vào tháng 2 năm 2007.

Hệ thống tình báo thống nhất Amalga của Microsoft - “Microsoft Expands Presence in Healthcare IT Industry with Acquisition of Health Intelligence Software Azyxxi”, Microsoft press release, July 26, 2006 (http://www.microsoft.com/en-us/news/press/2006/jul06/07-26azyxxiacquisitionpr.aspx). Dịch vụ Amalga bây giờ là một phần trong liên doanh của Microsoft với General Electric, gọi là Caradigm.

Amazon và “hợp tác lọc” - IPO Prospectus, May 1997 (http://
Amazon và “hợp tác lọc” - IPO Prospectus, May 1997 (http://
000868.txt).

Các bộ vi xử lý của xe hơi - Nick Valery, “Tech.View: Cars and Software Bugs”, *Economist.com,* May 16, 2010 (http://www.economist.com/blogs/babbage/2010/05/techview_cars_and_software_bugs).

Maury gọi các tàu là “đài quan sát nổi” - Maury, *The Physical Geography of the Sea.*

Về Viện Chi phí chăm sóc sức khỏe - Sarah Kliff, “A Database That Could Revolutionize Health Care”, *Washington Post,* May 21, 2012.

Google và thỏa thuận ITA - Claire Cain Miller, “U.S. Clears Google Acquisition of Travel Software”, *New York Times,* April 8, 2011 (http://www.nytimes.com/2011/04/09/technology/09google.html?_r=0).

Đối thoại từ bộ phim *Moneyball*, đạo diễn Bennett Miller,

Columbia Pictures, 2011.

Về phòng vé Hollywood so với doanh số bán trò chơi điện tử - Đối với phim, xem Brooks Barnes, "A Year of Disappointment at the Movie Box Office", *New York Times,* December 25, 2011 (http://www.nytimes.com/2011/12/26/business/media/a-year-ofdisappointment-for-hollywood.html). Đối với trò chơi điện tử, xem "Factbox: A Look at the $65 billion Video Games Industry", Reuters, June 6, 2011 (http://uk.reuters.com/article/2011/06/06/ us-videogames-factbox-idUKTRE75552I20110606).

Phân tích dữ liệu Zynga Nick Wingfield, "Virtual Products, Real Profits: Players Spend on Zynga's Games, but Quality Turns Some Off", *Wall Street Journal,* September 9, 2011 (http://online.wsj.com/article/SB10001424053111904823804576502442835446.html).

Trích dẫn của Ken Rudin - Erik Schlie, Jörg Rheinboldt, and Niko Waesche, *Simply Seven: Seven Ways to Create a Sustainable Internet Business* (Palgrave Macmillan, 2011). p. 7.

Trích dẫn của Auden - W. H. Auden, "For the Time Being", 1944.

Nghiên cứu Brynjolfsson - Erik Brynjolfsson, Lorin Hitt, and Heekyung Kim, "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" working paper, April 2011 (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486).

Về Rolls-Royce - "Rolls-Royce: Britain's Lonely High-Flier", *The Economist,* January 8, 2009 (http://www.economist.com/node/12887368). Figures updated from press office, November 2012.

Erik Brynjolfsson, Andrew McAfee, Michael Sorell, and Feng Zhu,

“Scale Without Mass: Business Process Replication and Industry Dynamics”, Harvard Business School working paper, September 2006 (http://www.hbs.edu/research/pdf/07-016.pdf also http:// hbswk.hbs.edu/item/5532.html).

Về chuyển biến hướng sang các chủ sở hữu dữ liệu ngày càng lớn - Yannis Bakos and Erik Brynjolfsson, “Bundling Information Goods: Pricing, Profits, and Efficiency”, *Management Science* 45 (December 1999), pp. 1613-30.

8. NHỮNG RỦI RO

Về Stasi - Rất tiếc là phần lớn các tài liệu đều bằng tiếng Đức, ngoại trừ một nghiên cứu rất hay là Kristie Macrakis, *Seduced by Secrets: Inside the Stasi’s Spy-Tech World* (Cambridge University Press, 2008). Chúng tôi cũng giới thiệu bộ phim đoạt giải Oscar *The Lives of Others*, do Florian Henckel von Donnersmark đạo diễn, Buena Vista / Sony Pictures năm 2006.

Camera giám sát gần nhà của Orwell - “George Orwell, Big Brother Is Watching Your House”, *The Evening Standard,* March 31, 2007 (http://www.thisislondon.co.uk/news/george-orwellbig-brother-is-watching-your-house-7086271.html).

Về Equifax và Experian - Daniel J. Solove, *The Digital Person: Technology and Privacy in the Information Age* (NYU Press, 2004), pp. 20-21.

Về địa chỉ khu phố của người Nhật Bản tại Washington được trao cho nhà chức trách Mỹ - J. R. Minkel, “The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II”, *Scientific American,* March 30, 2007 (http://www.scientificamerican.com/ article.cfm?id=confirmed-the-us-census-b).

Về dữ liệu được sử dụng bởi Đức quốc xã ở Hà Lan - William Seltzer and Margo Anderson, "The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses", *Social Research* 68 (2001), pp. 481-513.

Về IBM và Holocaust - Edwin Black, *IBM and the Holocaust* (Crown, 2003).

Về số lượng dữ liệu do các đồng hồ thông minh thu thập - Elias Leake Quinn, "Smart Metering and Privacy: Existing Law and Competing Policies; A Report for the Colorado Public Utility Commission", Spring 2009 (http://www.w4ar.com/Danger_of_Smart_Meters_Colorado_Report.pdf); Joel M. Margolis, "When Smart Grids Grow Smart Enough to Solve Crimes", Neustar, March 18, 2010 (http://energy.gov/sites/prod/files/gcprod/documents/Neustar_Comments_DataExhibitA.pdf)

Tài liệu của Fred Cate về xin phép và cho phép - Fred H. Cate, "The Failure of Fair Information Practice Principles", in Jane K. Winn, ed., *Consumer Protection in the Age of the "Information Economy"* (Ashgate, 2006), p. 341 et seq.

Về phát hành dữ liệu AOL - Michael Barbaro and Tom Zeller Jr., "A Face Is Exposed for AOL Searcher No. 4417749", *New York Times,* August 9, 2006; Matthew Karnitschnig and Mylene Mangalindan, "AOL Fires Technology Chief After Web-Search Data Scandal", *Wall Street Journal,* August 21, 2006.

Netflix xác định cá nhân - Ryan Singel, "Netflix Spilled Your *Brokeback Mountain* Secret, Lawsuit Claims", *Wired,* December 17, 2009 (http://www.wired.com/threatlevel/2009/12/netflixprivacy-lawsuit/).

Về việc phát hành dữ liệu Netflix - Arvind Narayanan and Vitaly Shmatikov, "Robust De-Anonymization of Large Sparse

Datasets", *Proceedings of the 2008 IEEE Symposium on Security and Privacy,* p. 111 et seq. (http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf); Arvind Narayanan and Vitaly Shmatikov, "How to Break the Anonymity of the Netflix Prize Dataset", October 18, 2006, arXiv:cs/0610105 [cs.CR] (http://arxiv.org/abs/cs/0610105).

Về việc xác định cá nhân từ ba đặc tính - Philippe Golle, "Revisiting the Uniqueness of Simple Demographics in the US Population", *Association for Computing Machinery Workshop on Privacy in Electronic Society* 5 (2006), p. 77.

Về sự suy yếu cấu trúc của ẩn danh hóa - Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", 57 *UCLA Law Review* 1701 (2010).

Về sự ẩn danh của đồ thị xã hội - Lars Backstrom, Cynthia Dwork, and Jon Kleinberg, "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography", *Communications of the Association of Computing Machinery,* December 2011, p. 133.

Các "hộp đen" của xe hơi - "Vehicle Data Recorders: Watching Your Driving", *The Economist,* June 23, 2012 (http://www.economist.com/node/21557309).

Thu thập dữ liệu NSA - Dana Priest and William Arkin, "A Hidden World, Growing Beyond Control", *Washington Post,* July 19, 2010 (http://projects.washingtonpost.com/top-secretamerica/articles/a-hidden-world-growing-beyond-control/ print/). Juan Gonzalez, "Whistleblower: The NSA Is Lying - U.S. Government Has Copies of Most of Your Emails", *Democracy Now,* April 20, 2012 (http://www.democracynow.org/2012/4/20/whistleblower_the_nsa_is_ lying_us). William Binney, "Sworn

Declaration in the Case of Jewel v. NSA", filed July 2, 2012 (http:// publicintelligence.net/binney-nsa-declaration/).

Việc giám sát đã thay đổi thế nào với dữ liệu lớn - Patrick Radden Keefe, "Can Network Theory Thwart Terrorists?" *New York Times,* March 12, 2006 (http://www.nytimes.com/2006/03/12/ magazine/312wwln_essay.html).

Đối thoại trong phim *Minority Report* của đạo diễn Steven Spielberg, DreamWorks / 20th Century Fox, 2002. Cuộc đối thoại chúng tôi trích dẫn là tóm tắt rất gọn. Bộ phim dựa trên một truyện ngắn năm 1958 của Philip K. Dick, nhưng có những sự khác biệt đáng kể giữa hai phiên bản. Cụ thể, cảnh mở đầu về người chồng bị cắm sừng không xuất hiện trong cuốn sách, và câu hỏi triết lý hóc búa về tiền tội phạm được trình bày trong phim của Spielberg đầy đủ hơn trong truyện. Do đó, chúng tôi đã chọn mô tả sự tương đồng so với bộ phim.

Câc thí dụ về giám sát tiên đoán - James Vlahos, "The Department Of Pre-Crime", *Scientific American* 306 (January 2012), pp. 62-67.

Về Future Attribute Screening Technology (FAST) - Sharon Weinberger, "Terrorist 'Pre-crime' Detector Field Tested in United States", *Nature,* May 27, 2011 (http://www.nature. com/news/2011/110527/full/news.2011.323.html); Sharon Weinberger, "Intentto Deceive", *Nature* 465 (May 2010), pp. 412-415. Về vấn đề dương tính giả - Alexander Furnas, "Homeland Security's 'Pre-Crime' Screening Will Never Work", *The Atlantic Online,* April 17, 2012 (http://www.theatlantic.com/technology/ archive/2012/04/homeland-securitys-pre-crime-screeningwill-never-work/255971/).

Về điểm của học sinh và phí bảo hiểm - Tim Query, "Grade

Inflation and the Good-Student Discount, *Contingencies Magazine*, American Academy of Actuaries, May-June 2007 (http://www.contingencies.org/mayjun07/tradecraft.pdf). Về những nguy hiểm của lập hồ sơ - Bernard E. Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age* (University of Chicago Press, 2006).

Về công trình của Richard Berk - Richard Berk, "The Role of Race in Forecasts of Violent Crime", *Race and Social Problems* 1 (2009), pp. 231-242.

Về sự say mê dữ liệu của McNamara - Phil Rosenzweig, "Robert S. McNamara and the Evolution of Modern Management", *Harvard Business Review,* December 2010 (http://hbr.org/2010/12/robert-s-mcnamara-and-the-evolution-ofmodern-management/ar/pr).

Về thành công của "Những đứa trẻ thần đồng" trong Thế chiến II - John Byrne, *The Whiz Kids* (Doubleday, 1993).

Về McNamara tại Ford - David Halberstam, *The Reckoning* (William Morrow, 1986), pp. 222-245.

Cuốn sách của Kinnard - Douglas Kinnard, *The War Managers* (University Press of New England, 1977), pp. 71-25.

Về câu trích dẫn "Chúng ta tin ở Chúa - còn tất cả những thứ khác thì mang đến dữ liệu" - Câu trích này thường được gán cho W. Edwards Deming.

Về Ted Kennedy và danh sách cấm bay - Sara Kehaulani Goo, "Sen. Kennedy Flagged by No-Fly List", *Washington Post,* August 20, 2004, p. A01 (http://www.washingtonpost.com/wp-dyn/articles/A17073-2004Aug19.html).

Biện pháp tuyển dụng của Google - Xem Douglas Edwards, *I'm*

Feeling Lucky: The Confessions of Google Employee Number 59 (Houghton Mifflin Harcourt, 2011), p. 9; Steven Levy, *In the Plex* (Simon and Schuster, 2011), pp. 140-141. Trớ trêu thay, người đồng sáng lập của Google từng muốn thuê Steve Jobs làm CEO (mặc dù ông không có bằng đại học); Levy, p. 80.

Thử nghiệm 41 tỷ lệ chiết giảm của màu xanh lam - Laura M. Holson, "Putting a Bolder Face on Google", *New York Times,* March 1, 2009 (http://www.nytimes.com/2009/03/01/business/01marissa.html).

Giám đốc thiết kế của Google từ chức - Doug Bowman, "Goodbye, Google", blog post, March 20, 2009 (http://stopdesign.com/archive/2009/03/20/goodbye-google.html).

Trích dẫn của Jobs - Steve Lohr, "Can Apple Find More Hits Without Its Tastemaker?" *New York Times,* January 18, 2011, p. B1 (http://www.nytimes.com/2011/01/19/technology/companies/19innovate.html).

Cuốn sách của Scott - James Scott, *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (Yale University Press, 1998).

Trích dẫn của McNamara từ năm 1967 - Từ bài nói chuyện tại Millsaps College ở Jackson, Mississippi, được trích dẫn trong *Harvard Business Re*view, tháng 12 năm 2010.

Về lời biện hộ của McNamara - Robert S. McNamara with Brian VanDeMark, *In Retrospect: The Tragedy and Lessons of Vietnam* (Random House, 1995), pp. 48, 270.

9. KIỂM SOÁT

Về việc sưu tập sách thư viện của Đại học Cambridge - Marc Drogin, *Anathema! Medieval Scribes and the History of Book Curses*

(Allanheld and Schram, 1983), p. 37.

Về trách nhiệm giải trình và sự riêng tư - Trung tâm Quản lý chính sách thông tin đã tham gia trong một dự án kéo dài nhiều năm về những nét chung trong trách nhiệm giải trình và sự riêng tư, xem http://www.informationpolicycentre.com/accountability-based_privacy_governance/.

Về ngày hết hạn của dữ liệu - Mayer-Schönberger, *Delete.* "Differential privacy" - Cynthia Dwork, "A Firm Foundation for Private Data Analysis", *Communications of the ACM,* January 2011, pp. 86-95.

Facebookvà quyền riêng tư khác biệt - A. Chin and A. Klinefelter, "Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study", 90 *North Carolina Law Review* 1417 (2012); A. Haeberlen et al., "Differential Privacy Under Fire", http://www.cis.upenn.edu/~ahae/papers/fuzzsec2011.pdf.

Các công ty bị nghi ngờ thông đồng - Đã có nghiên cứu trong lĩnh vực này, xem Pim Heijnen, Marco A. Haan, and Adriaan R. Soetevent, "Screening for Collusion: A Spatial Statistics Approach", Discussion Paper TI 2012-058/1, Tinbergen Institute, The Netherlands, 2012 (http://www.tinbergen.nl/discussionpapers/12058.pdf).

Về các đại diện bảo vệ dữ liệu của công ty Đức - Viktor MayerSchönberger, "Beyond Privacy, Beyond Rights: Towards a 'Systems' Theory of Information Governance", 98 *California Law Review* 1853 (2010).

Về khả năng tương tác - John Palfrey and Urs Gasser, *Interop: The Promise and Perils of Highly Interconnected Systems* (Basic Books, 2012).

10. TIẾP THEO

Thông tin về Mike Flowers và các phân tích của thành phố New York được dựa trên cuộc phỏng vấn với Cukier vào tháng 7 năm 2012; xem Alex Howard, “Predictive data analytics is saving lives and taxpayer dollars in New York City”, *O’Reilly Media,* June 26, 2012 (http://strata.oreilly.com/2012/06/predictive-dataanalytics-big-data-nyc.html).

Về Walmart và Pop-Tarts - Hays, “What Wal-Mart Knows About Customers’ Habits”.

Ứng dụng của dữ liệu lớn trong các khu ổ chuột và trong mô hình hóa những phong trào tị nạn - Nathan Eagle, “Big Data, Global Development, and Complex Systems”, http://www.youtube. com/watch?v=yaivtqlu7iM.

Nhận thức về thời gian - Benedict Anderson, *Imagined Communities* (Verso, 2006).

“Quá khứ là khúc dạo đầu” - William Shakespeare, “The Tempest”, Act 2, Scene I.

Hệ thống máy tính của Apollo 11 - David A. Mindell, *Digital Apollo: Human and Machine in Spaceflight* (MIT Press, 2008).

TÀI LIỆU THAM KHẢO

Alter, Alexandra. "Your E-Book Is Reading You". *Wall Street Journal,* June 29, 2012 (http://online.wsj.com/article/SB10001424052702 304870304577490950051438304.html).

Anderson, Benedict. *Imagined Communities,* New Edition. Verso, 2006. Anderson, Chris. "The End of Theory". *Wired* 16, issue 7 (July 2008 (http://www.wired.com/science/discoveries/ magazine/16-07/pb_theory).

Asur, Sitaram, and Bernardo A. Huberman. "Predicting the Future with Social Media". *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,* pp. 492-499. (An online version is available at http://www.hpl.hp.com/research/scl/papers/socialmedia/ socialmedia.pdf.)

Asur, Sitaram, and Bernardo A. Huberman. "Predicting the Future with Social Media". *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,* pp. 492-499. (Một phiên bản trực tuyến có sẵn tại http://www.hpl.hp.com/research/scl/papers/socialmedia / socialmedia.pdf.)

Ayres, Ian. Super Crunchers: Why Thinking-By-Numbers Is the New Way to Be Smart. *Bantam Dell, 2007.*

Babbie, Earl. Practice of Social Research, *12th ed. 2010.*

Backstrom, Lars, Cynthia Dwork, and Jon Kleinberg. "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden

Patterns, and Structural Steganography". *Communications of the ACM,* December 2011, pp. 133-141.

Bakos, Yannis, and Erik Brynjolfsson. "Bundling Information Goods: Pricing, Profits, and Efficiency". *Management Science* 45 (December 1999), pp. 1613-30.

Banko, Michele, and Eric Brill. "Scaling to Very Very Large Corpora for Natural Language Disambiguation". Microsoft Research, 2001, p. 3 (http://acl.ldc.upenn.edu/P/P01/P01-1005.pdf).

Barbaro, Michael, and Tom Zeller Jr. "A Face Is Exposed for AOL Searcher No. 4417749". *New York Times,* August 9, 2006 (http://www.nytimes.com/2006/08/09/technology/09aol.html).

Barbaro, Michael, và Tom Zeller Jr " Một khuôn mặt là xúc cho AOL Searcher Barnes, Brooks. "A Year of Disappointment at the Movie Box Office", *New York Times,* December 25, 2011 (http://www.nytimes.com/2011/12/26/business/media/a-year-ofdisappointment-for-hollywood.html).

Barnes, Brooks. "A Year of Disappointment at the Movie Box Office", *New York Times,* December 25, 2011 (http://www.nytimes. com/2011/12/26/business/media/a-year-of-disappointmentfor-hollywood.html).

Beaty, Janice. *Seeker of Seaways: A Life of Matthew Fontaine Maury, Pioneer Oceanographer.* Pantheon Books, 1966.

Berger, Adam L., et al. "The Candide System for Machine Translation". *Proceedings of the 1994 ARPA Workshop on Human Language Technology* (1994) (http://aclweb.org/anthology-new/H/H94/ H94-1100.pdf).

Berk, Richard. "The Role of Race in Forecasts of Violent Crime".

Race and Social Problems 1 (2009), pp. 231-242.

Black, Edwin. *IBM and the Holocaust.* Crown, 2003.

boyd, danah, and Kate Crawford. "Six Provocations for Big Data". Research paper presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society", September 21, 2011 (http://ssrn.com/abstract=1926431).

Brown, Brad, Michael Chui, and James Manyika. "Are You Ready for the Era of 'Big Data'?" *McKinsey Quarterly,* October 2011, p. 10.

Brynjolfsson, Erik, Andrew McAfee, Michael Sorell, and Feng Zhu. "Scale Without Mass: Business Process Replication and Industry Dynamics". HBS working paper, September 2006 (http://www. hbs.edu/research/pdf/07-016.pdf; also http://hbswk.hbs.edu/ item/5532.html).

Brynjolfsson, Erik, Lorin Hitt, and Heekyung Kim. "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" *ICIS 2011 Proceedings,* Paper 13 (http://aisel. aisnet.org/icis2011/proceedings/economicvalueIS/13; also available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486).

Byrne, John. *The Whiz Kids.* Doubleday, 1993.

Cate, Fred H. "The Failure of Fair Information Practice Principles". In Jane K. Winn, ed., *Consumer Protection in the Age of the "Information Economy"* (Ashgate, 2006), p. 341 et seq.

Chin, A., and A. Klinefelter. "Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study". 90 *North Carolina Law Review* 1417 (2012).

Crosby, Alfred. *The Measure of Reality: Quantification and Western Society, 1250-1600.* Cambridge University Press, 1997.

Cukier, Kenneth. “Data, Data Everywhere”. *The Economist* Special Report, February 27, 2010, pp. 1-14.

___. “Tracking Social Media: The Mood of the Market”. *Economist.com,* June 28, 2012 (http://www.economist.com/blogs/graphicdetail/2012/06/tracking-social-media).

Davenport, Thomas H., Paul Barth, and Randy Bean. “How ‘Big Data’ Is Different”. *Sloan Review,* July 30, 2012 (http://sloanreview. mit.edu/the-magazine/2012-fall/54104/how-big-data-isdifferent/).

Di Quinzio, Melanie, and Anne McCarthy. “Rabies Risk Among Travellers”. *CMAJ* 178, no. 5 (2008), p. 567.

Drogin, Marc. *Anathema! Medieval Scribes and the History of Book Curses.* Allanheld and Schram, 1983.

Dugas, A. F., et al. “Google Flu Trends: Correlation with Emergency Department Influenza Rates and Crowding Metrics”. CID Advanced Access, January 8, 2012. DOI 10.1093/cid/cir883.

Duggan, Mark, and Steven D. Levitt. “Winning Isn’t Everything: Corruption in Sumo Wrestling”. *American Economic Review* 92 (2002), pp. 1594-1605 (http://pricetheory.uchicago.edu/levitt/Papers/DugganLevitt2002.pdf).

Duhigg, Charles. The Power of Habit: Why We Do What We Do in Life and Business. *Random House, 2012.*

Duhigg, Charles. “How Companies Learn Your Secrets”. *New York Times,* February 16, 2012

(http://www.nytimes.com/2012/02/19/magazine/shoppinghabits.html).

Dwork, Cynthia. “A Firm Foundation for Private Data Analysis”. *Communications of the ACM,* January 2011, pp. 86-95 (http://dl.acm.org/citation.cfm?id=1866739.1866758).

Economist, The. “Rolls-Royce: Britain’s Lonely High-Flier”. *The Economist,* January 8, 2009 (http://www.economist.com/node/12887368).

___. “Building with Big Data: The Data Revolution Is Changing the Landscape of Business”. *The Economist,* May 26, 2011 (http:// www.economist.com/node/18741392/).

___. “Official Statistics: Don’t Lie to Me, Argentina”. *The Economist,* February 25, 2012 (http://www.economist.com/node/21548242).

___. “Counting Every Moment”. *The Economist,* March 3, 2012 (http://www.economist.com/node/21548493).

___. “Vehicle Data Recorders: Watching Your Driving”. *The Economist,* June 23, 2012 (http://www.economist.com/node/21557309).

Edwards, Douglas. *I’m Feeling Lucky: The Confessions of Google Employee Number 59.* Houghton Mifflin Harcourt, 2011.

Ehrenberg, Rachel. “Predicting the Next Deadly Manhole Explosion”. *Wired,* July 7, 2010 (http://www.wired.com/wiredscience/2010/07/manholeexplosions).

Eisenstein, Elizabeth L. *The Printing Revolution in Early Modern Europe.* Cambridge University Press, 1993.

Etzioni, Oren, C. A. Knoblock, R. Tuchinda, and A. Yates. “To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price”. SIGKDD ’03, August 24-27, 2003 (http://knight.cis.temple.edu/~yates//papers/hamlet-kdd03.pdf).

Frei, Patrizia, et al. “Use of Mobile Phones and Risk of Brain Tumours: Update of Danish Cohort Study”. *BMJ* 2011, 343 (http://www. bmj.com/content/343/bmj.d6387).

Furnas, Alexander. “Homeland Security’s ‘Pre-Crime’ Screening Will Never Work”. *The Atlantic Online,* April 17, 2012 (http://www.theatlantic.com/technology/archive/2012/04/homelands pre-crime-screeningwill-never-work/255971/). Garton Ash, Timothy. *The File.* Atlantic Books, 2008.

Geron, Tomio. “Twitter’s Dick Costolo: Twitter Mobile Ad Revenue Beats Desktop on Some Days”. *Forbes,* June 6, 2012 (http://
www.forbes.com/sites/tomiogeron/2012/06/06/twitters-dickcostolo-mobile-ad-revenue-beats-desktop-on-some-days/).

Ginsburg, Jeremy, et al. “Detecting Influenza Epidemics Using Search Engine Query Data”. *Nature* 457 (2009), pp. 1012-14 (http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html).

Golder, Scott A., and Michael W. Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures”. *Science* 333 (September 30, 2011), pp. 1878-81.

Golle, Philippe. “Revisiting the Uniqueness of Simple Demographics in the US Population”. *Association for Computing Machinery Workshop on Privacy in Electronic Society* 5 (2006), pp. 77-80.

Goo, Sara Kehaulani. “Sen. Kennedy Flagged by No-Fly List”.

Washington Post, August 20, 2004, p. A01 (http://www.washingtonpost.com/wp-dyn/articles/A17073-2004Aug19.html).

Haeberlen, A., et al. “Differential Privacy Under Fire”. In *SEC’11: Proceedings of the 20th USENIX conference on Security,* p. 33 (http://www.cis.upenn.edu/~ahae/papers/fuzz-sec2011.pdf).

Halberstam, David. *The Reckoning.* William Morrow, 1986.

Haldane, J. B. S. “On Being the Right Size”. *Harper’s Magazine,* March 1926 (http://harpers.org/archive/1926/03/on-being-the-rightsize/).

Halevy, Alon, Peter Norvig, and Fernando Pereira. “The Unreasonable Effectiveness of Data”. *IEEE Intelligent Systems,* March/April 2009, pp. 8-12.

Harcourt, Bernard E. Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age. *University of Chicago Press, 2006*.

Hardy, Quentin. “Bizarre Insights from Big Data”. *NYTimes.com,* March 28, 2012 (http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/).

Hays, Constance L. “What Wal-Mart Knows About Customers’ Habits”. *New York Times,* November 14, 2004 (http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html).

Hearn, Chester G. Tracks in the Sea: Matthew Fontaine Maury and the Mapping of the Oceans. *International Marine/McGraw-Hill, 2002.*

Helland, Pat. “If You Have Too Much Data then “ ‘Good Enough’ Is Good Enough”. *Communications of the ACM,* June 2011, p. 40

et seq.

Hilbert, Martin, and Priscilla López. “The World’s Technological Capacity to Store, Communicate, and Compute Information”. *Science* 1 (April 2011), pp. 60-65.

___. “How to Measure the World’s Technological Capacity to Communicate, Store and Compute Information?” *International Journal of Communication* (2012), pp. 1042-55 (ijoc.org/ojs/index.php/ijoc/article/viewFile/1562/742).

Holson, Laura M. “Putting a Bolder Face on Google”. *New York Times,* March 1, 2009, p. BU 1 (http://www.nytimes.com/2009/03/01/business/01marissa.html).

Hopkins, Brian, and Boris Evelson. “Expand Your Digital Horizon with Big Data”. Forrester, September 30, 2011.

Hotz, Robert Lee. “The Really Smart Phone”. *Wall Street Journal,* April 22, 2011 (http://online.wsj.com/article/SB100014240527487045476045

Hutchins, John. “The First Public Demonstration of Machine Translation: The Georgetown-IBM System, 7th January 1954”. November 2005 (http://www.hutchinsweb.me.uk/GUIBM-2005.pdf).

Inglehart, R., and H. D. Klingemann. *Genes, Culture and Happiness.* MIT Press, 2000.

Isaacson, Walter. *Steve Jobs.* Simon and Schuster, 2011.

Kahneman, Daniel. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, 2011.

Kaplan, Robert S., and David P. Norton. *Strategy Maps: Converting Intangible Assets into Tangible Outcomes.* Harvard Business Review Press, 2004.

Karnitschnig, Matthew, and Mylene Mangalindan. “AOL Fires Technology Chief After Web-Search Data Scandal”. *Wall Street Journal,* August 21, 2006.

Keefe, Patrick Radden. “Can Network Theory Thwart Terrorists?” *New York Times,* March 12, 2006 (http://www.nytimes.com/2006/03/12/magazine/312wwln_essay.html).

Kinnard, Douglas. *The War Managers.* University Press of New England, 1977.

Kirwan, Peter. “This Car Drives Itself”. *Wired UK,* January 2012 (http://www.wired.co.uk/magazine/archive/2012/01/features/thiscar-drives-itself).

Kliff, Sarah. “A Database That Could Revolutionize Health Care”. *Washington Post,* May 21, 2012.

Kruskal, William, and Frederick Mosteller. “Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939”. *International Statistical Review* 48 (1980), pp. 169-195.

Laney, Doug. “To Facebook You’re Worth $80.95”. *Wall Street Journal,* May 3, 2012 (http://blogs.wsj.com/cio/2012/05/03/to-facebookyoure-worth-80-95/).

Latour, Bruno. *The Pasteurization of France.* Harvard University Press, 1993.

Levitt, Steven D., and Stephen J. Dubner. *Freakonomics: A Rogue*

Economist Explores the Hidden Side of Everything. William Morrow, 2009.

Levy, Steven. *In the Plex.* Simon and Schuster, 2011.

Lewis, Charles Lee. *Matthew Fontaine Maury: The Pathfinder of the Seas.* U.S. Naval Institute, 1927.

Lohr, Steve. "Can Apple Find More Hits Without Its Tastemaker?" *New York Times,* January 18, 2011, p. B1 (http://www.nytimes.com/2011/01/19/technology/companies/19innovate.html).

Lowrey, Annie. "Economists' Programs Are Beating U.S. at Tracking Inflation". *Washington Post,* December 25, 2010 (http://www.washingtonpost.com/wp-dyn/content/article/2010/12/25/AR2010122502600.html).

Macrakis, Kristie. *Seduced by Secrets: Inside the Stasi's Spy-Tech World.* Cambridge University Press, 2008.

Manyika, James, et al. "Big Data: The Next Frontier for Innovation, Competition, and Productivity". *McKinsey Global Institute,* May 2011 (http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation).

Marcus, James. *Amazonia: Five Years at the Epicenter of the Dot.Com Juggernaut.* The New Press, 2004.

Margolis, Joel M. "When Smart Grids Grow Smart Enough to Solve Crimes". Neustar, March 18, 2010 (http://energy.gov/sites/prod/files/gcprod/documents/Neustar_Comments_DataExhibitA.pdf).

Maury, Matthew Fontaine. *The Physical Geography of the Sea.* Harper, 1855.

Mayer-Schönberger, Viktor. “Beyond Privacy, Beyond Rights: Towards a ‘Systems’ Theory of Information Governance”. 98 *California Law Review* 1853 (2010).

___. *Delete: The Virtue of Forgetting in the Digital Age.* Princeton University Press, 2nd ed., 2011.

McGregor, Carolyn, Christina Catley, Andrew James, and James Padbury. “Next Generation Neonatal Health Informatics with Artemis”. In European Federation for Medical Informatics, *User Centred Networked Health Care,* ed. A. Moen et al. (IOS Press, 2011), p. 117 et seq.

McNamara, Robert S., with Brian VanDeMark. *In Retrospect: The Tragedy and Lessons of Vietnam.* Random House, 1995.

Mehta, Abhishek. “Big Data: Powering the Next Industrial Revolution”. Tableau Software White Paper, 2011.

Michel, Jean-Baptiste, et al. “Quantitative Analysis of Culture Using Millions of Digitized Books”. *Science* 331 (January 14, 2011), pp. 176-182 (http://www.sciencemag.org/content/331/6014/176.abstract).

Miller, Claire Cain. “U.S. Clears Google Acquisition of Travel Software”. *New York Times,* April 8, 2011 (http://www.nytimes.com/2011/04/09/technology/09google.html?_r=0).

Mills, Howard. “Analytics: Turning Data into Dollars”. *Forward Focus,* December 2011 (http://www.deloitte.com/assets/ Dcom-UnitedStates/Local%20Assets/Documents/FSI/US_FSI_Forward%20Focus_Analytics_Turning%20data%20into%20

dollars_120711.pdf).

Mindell, David A. Digital Apollo: Human and Machine in Spaceflight. *MIT Press, 2008.*

Minkel, J. R. "The U.S. Census Bureau Gave Up Names of JapaneseAmer-icans in WW II". *Scientific American,* March 30, 2007 (http://www.scientificamerican.com/article.cfm?id=confirmedthe-us-census-b).

Murray, Alexander. *Reason and Society in the Middle Ages.* Oxford University Press, 1978.

Nalimov, E. V., G. McC. Haworth, and E. A. Heinz. "Space-Efficient Indexing of Chess Endgame Tables". *ICGA Journal* 23, no. 3 (2000), pp. 148-162.

Narayanan, Arvind, and Vitaly Shmatikov. "How to Break the Anonymity of the Netflix Prize Dataset". October 18, 2006, arXiv:cs/0610105 (http://arxiv.org/abs/cs/0610105).

___. "Robust De-Anonymization of Large Sparse Datasets". *Proceedings of the 2008 IEEE Symposium on Security and Privacy,* p. 111 (http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf).

Nazareth, Rita, and Julia Leite. "Stock Trading in U.S. Falls to Lowest Level Since 2008". *Bloomberg,* August 13, 2012 (http://www. bloomberg.com/news/2012-08-13/stock-trading-in-u-s-hitslowest-level-since-2008-as-vix-falls.html).

Negroponte, Nicholas. *Being Digital.* Alfred Knopf, 1995.

Neyman, Jerzy. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection". *Journal of the Royal Statistical Society* 97, no. 4 (1934), pp. 558-625.

Ohm, Paul. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization”. 57 *UCLA Law Review* 1701 (2010).

Onnela, J. P., et al. “Structure and Tie Strengths in Mobile Communication Networks”. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 104 (May 2007), pp. 7332-36 (http://nd.edu/~dddas/Papers/PNAS0610245104v1.pdf).

Palfrey, John, and Urs Gasser. *Interop: The Promise and Perils of Highly Interconnected Systems.* Basic Books, 2012.

Pearl, Judea. *Causality: Models, Reasoning and Inference,* 2nd ed. Cambridge University Press, 2009.

President’s Council of Advisors on Science and Technology. “Report to the President and Congress, Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology”. December 2010 (http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcastnitrd-report-2010.pdf).

Priest, Dana and William Arkin. “A Hidden World, Growing Beyond Control”. *Washington Post,* July 19, 2010 (http://projects.washingtonpost.com/top-secret-america/articles/a-hiddenworld-growing-beyond-control/print/).

Query, Tim. “Grade Inflation and the Good-Student Discount”. *Contingencies Magazine,* American Academy of Actuaries, MayJune 2007 (http://www.contingencies.org/mayjun07/tradecraft. pdf).

Quinn, Elias Leake. “Smart Metering and Privacy: Existing Law and Competing Policies; A Report for the Colorado Public Utility Commission”. Spring 2009 (http://www.w4ar.com/Danger_of_Smart_Meters_Colorado_Report.pdf).

Reshef, David, et al. “DetectingNovel Associations in Large Data Sets”. *Science* (2011), pp. 1518-24.

Rosenthal, Jonathan. “Banking Special Report”. *The Economist,* May 19, 2012, pp. 7-8.

Rosenzweig, Phil. “Robert S. McNamara and the Evolution of Modern Management”. *Harvard Business Review,* December 2010, pp. 87-93 (http://hbr.org/2010/12/robert-s-mcnamara-and-theevolution-of-modern-management/ar/pr).

Rudin, Cynthia, et al. “21st-Century Data Miners Meet 19th-Century Electrical Cables”. *Computer,* June 2011, pp. 103-105.

___. “Machine Learning for the New York City Power Grid”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.2 (2012), pp. 328-345 (http://hdl.handle. net/1721.1/68634).

Rys, Michael. “Scalable SQL”. *Communications of the ACM,* June 2011, 48, pp. 48-53.

Salathé, Marcel, and Shashank Khandelwal. “Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control”. *PlOS Computational Biology* 7, no. 10 (October 2011).

Savage, Mike, and Roger Burrows. “The Coming Crisis of Empirical Sociology”. *Sociology* 41 (2007), pp. 885-899.

Schlie, Erik, Jörg Rheinboldt, and Niko Waesche. *Simply Seven: Seven Ways to Create a Sustainable Internet Business.* Palgrave

Macmillan, 2011.

Scanlon, Jessie. "Luis von Ahn: The Pioneer of 'Human Computation.' " *Businessweek,* November 3, 2008 (http://www.businessweek. com/stories/2008-11-03/ luis-von-ahn-the-pioneer-of-humancomputation-businessweek-business-news-stock-market-andfinancial-advice).

Scism, Leslie, and Mark Maremont. "Inside Deloitte's Life-Insurance Assessment Technology". *Wall Street Journal,* November 19, 2010 (http://online.wsj.com/article/SB10001424052748704104104575622531084755588.html).

___. "Insurers Test Data Profiles to Identify Risky Clients". *Wall Street Journal,* November 19, 2010 (http://online.wsj.com/article/SB10001424052748704648604575620750998072986.html).

Scott, James. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed.* Yale University Press, 1998.

Seltzer, William, and Margo Anderson. "The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses". *Social Research* 68 (2001) pp. 481-513.

Silver, Nate. *The Signal and the Noise: Why So Many Predictions Fail - But Some Don't.* Penguin, 2012.

Singel, Ryan. "Netflix Spilled Your *Brokeback Mountain* Secret, Lawsuit Claims". *Wired,* December 17, 2009 (http://www.wired. com/threatlevel/2009/12/netflix-privacy-lawsuit/).

Smith, Adam. *The Wealth of Nations* (1776). Reprinted Bantam Classics, 2003. A free electronic version is available (http://www2.hn.psu. edu/faculty/jmanis/adam-

smith/Wealth-Nations.pdf).

Solove, Daniel J. *The Digital Person: Technology and Privacy in the Information Age.* NYU Press, 2004.

Surowiecki, James. “A Billion Prices Now”. *New Yorker,* May 30, 2011 (http://www.newyorker.com/talk/financial/2011/05/30/110530ta_talk_surowiecki).

Taleb, Nassim Nicholas. *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets.* Random House, 2008.

___. *The Black Swan: The Impact of the Highly Improbable.* 2nd ed., Random House, 2010.

Thompson, Clive. “For Certain Tasks, the Cortex Still Beats the CPU”. *Wired,* June 25, 2007 (http://www.wired.com/techbiz/it/magazine/15-7/ff_humancomp?currentPage=all).

Thurm, Scott. “Next Frontier in Credit Scores: Predicting Personal Behavior”. *Wall Street Journal,* October 27, 2011 (http://online.wsj.com/article/SB10001424052970203687504576655182086912.html).

Tsotsis, Alexia. “Twitter Is at 250 Million Tweets per Day, iOS 5 Integration Made Signups Increase 3x”. *TechCrunch,* October October million-tweets-per-day/).

Valery, Nick. “Tech.View: Cars and Software Bugs”. *The Economist,* May 16, 2010 (http://www.economist.com/blogs/babbage/2010/05/techview_cars_and_software_bugs).

Vlahos, James. “The Department Of Pre-Crime”. *Scientific American* 306 (January 2012), pp. 62-67.

Von Baeyer, Hans Christian. *Information: The New Language of Science.* Harvard University Press, 2005.

von Ahn, Luis, et al. "reCAPTCHA: Human-Based Character Recognition via Web Security Measures". *Science* 321 (September 12, 2008), pp. 1465-68 (http://www.sciencemag.org/content/321/5895/1465.abstract).

Watts, Duncan. Everything Is Obvious Once You Know the Answer: How Common Sense Fails Us. *Atlantic, 2011. Weinberger, David.* Everything Is Miscellaneous: The Power of the New Digital Disorder. *Times, 2007.*

Weinberger, Sharon. "Intent to Deceive". *Nature* 465 (May 2010), pp. 412-415 (http://www.nature.com/news/2010/100526/full/465412a.html).

___. "Terrorist 'Pre-crime' Detector Field Tested in United States". *Nature,* May 27, 2011 (http://www.nature.com/news/2011/110527/full/news.2011.323.html).

Whitehouse, David. "UK Science Shows Cave Art Developed Early". *BBC News Online,* October 3, 2001 (http://news.bbc.co.uk/1/hi/ sci/tech/1577421.stm).

Wigner, Eugene. "The Unreasonable Effectiveness of Mathematics in the Natural Sciences". *Communications on Pure and Applied Mathematics* 13, no. 1 (1960), pp. 1-14.

Wilks, Yorick. Machine Translation: Its Scope and Limits. *Springer, 2008.*

Wingfield, Nick. "Virtual Products, Real Profits: Players Spend on Zynga's Games, but Quality Turns Some Off". *Wall Street Journal,* September 9, 2011 (http://online.wsj.com/article/SB100014240

5311190482380457650244283541 3446.html).

LỜI CẢM ƠN

Cả hai chúng tôi đã may mắn được làm việc và học hỏi từ một cây đại thụ trong lĩnh vực mạng thông tin và đổi mới, Lewis M. Branscomb. Trí tuệ, tài hùng biện, năng lượng, tính chuyên nghiệp, sự hóm hỉnh, và óc tò mò vô tận của ông luôn tiếp tục truyền cảm hứng cho chúng tôi. Và với người đối tác tương đồng và khôn ngoan của ông, Connie Mullin, chúng tôi phải xin lỗi vì không lưu ý đến đề nghị của bà để đặt tên cuốn sách là “Siêu dữ liệu”.

Momin Malik là một trợ lý nghiên cứu tuyệt vời với trí tuệ và sự cần cù đặc biệt. Chúng tôi có hân hạnh được đại diện bởi Lisa Adams và David Miller của Tổ chức Garamond, một đại diện tuyệt vời trong mọi khía cạnh. Eamon Dolan, biên tập viên của chúng tôi, là đại diện cho lớp các biên tập viên quý hiếm, những người có cảm giác gần như hoàn hảo về việc làm thế nào để chỉnh sửa văn bản và thách thức suy nghĩ của chúng tôi, để kết quả tốt hơn nhiều so với chúng tôi có thể hy vọng. Chúng tôi cảm ơn tất cả mọi người tại Houghton Mifflin Harcourt, đặc biệt là Beth Burleigh Fuller và Ben Hyman. Ngoài ra còn có Camille Smith về việc biên tập bản thảo rất chuyên nghiệp của bà. Chúng tôi rất biết ơn James Fransham của *The Economist* về công việc kiểm tra tư liệu xuất sắc và những lời phê bình thông minh của ông đối với bản thảo.

Chúng tôi đặc biệt biết ơn tất cả những nhà chuyên môn dữ-liệu-lớn đã dành thời gian giải thích công việc của họ, đặc biệt là Oren Etzioni, Cynthia Rudin, Carolyn McGregor, và Mike Flowers.

***

Những lời cảm ơn cá nhân của Viktor: Tôi cảm ơn Philip Evans, người luôn luôn suy nghĩ trước hai bước và thể hiện ý tưởng của mình với độ chính xác và tài hùng biện, về các cuộc trao đổi kéo

dài hơn một thập kỷ. Tôi cũng biết ơn đồng nghiệp cũ David Lazer của tôi, một nhà hàn lâm dữ-liệu-lớn từ rất sớm và rất giỏi, mà rất nhiều lần tôi đã nhờ ông tư vấn.

Tôi cảm ơn những người tham gia Đối thoại Dữ liệu Kỹ thuật số Oxford 2011 (tập trung vào dữ liệu lớn), và đặc biệt là đồng chủ tịch Fred Cate, về các cuộc thảo luận vô cùng giá trị.

Viện Internet Oxford, nơi tôi làm việc, đã mang đến môi trường thuận lợi cho cuốn sách này, với rất nhiều đồng nghiệp của tôi tham gia vào nghiên cứu dữ-liệu-lớn. Tôi không thể nghĩ ra một nơi nào tốt hơn để viết nó. Tôi cũng muốn tỏ lòng biết ơn sự hỗ trợ của trường Keble College. Nếu không có sự hỗ trợ đó, tôi đã không được quyền truy cập vào một số trong những nguồn tham khảo quan trọng được sử dụng trong cuốn sách.

Gia đình luôn luôn phải chịu thiệt thòi lớn nhất khi có người viết một cuốn sách. Đó không chỉ là nhiều giờ tôi đã ngồi trước màn hình máy tính, vắng mặt để làm việc ở văn phòng, mà còn là nhiều, rất nhiều giờ tuy thân xác hiện hữu, nhưng lại bị chôn vùi trong suy nghĩ. Tôi cầu xin sự tha thứ từ vợ tôi Birgit và đứa con nhỏ Viktor của tôi. Tôi hứa sẽ cố gắng nhiều hơn.

Những lời cảm ơn cá nhân của Kenn: Tôi biết ơn nhiều các nhà khoa học dữ liệu lớn đã giúp đỡ, đặc biệt là Jeff Hammerbacher, Amr Awadallah, DJ Patil, Michael Driscoll, Michael Freed, và nhiều đồng nghiệp tại Google trong nhiều năm (bao gồm cả Hal Varian, Jeremy Ginsberg, Peter Norvig, và Udi Manber, cùng những người khác, và những cuộc trò chuyện ngắn vô giá với Eric Schmidt và Larry Page).

Suy nghĩ của tôi đã trở nên phong phú nhờ Tim O’Reilly, một nhà bác học của thời đại Internet, và bởi Marc Benioff của Salesforce.com, một người thầy. Những hiểu biết sâu sắc của Matthew Hindman luôn luôn là vô giá. James Guszcza của

Deloitte giúp ích cho tôi rất nhiều, và Geoff Hyatt, một người bạn cũ đang kinh doanh dữ liệu chuỗi cũng vậy. Xin gửi lời cảm ơn đặc biệt đến Pete Warden, vừa là một triết gia vừa là một nhà chuyên môn về dữ liệu lớn.

Nhiều bạn bè đã cung cấp những ý tưởng và tư vấn, bao gồm John Turner, Angelika Wolf, Niko Waesche, Katia Verresen, David Wishart, Anna Petherick, Blaine Harden và Jessica Kowal. Những người truyền cảm hứng cho các chủ đề trong cuốn sách bao gồm Blaise Aguera y Arcas, Eric Horvitz, David Auerbach, Gil Elbaz, Tyler Bell, Andrew Wyckoff và nhiều người khác tại OECD (Tổ chức Hợp tác Kinh tế và Phát triển), Stephen Brobst và đội ngũ tại Teradata, Anthony Goldbloom và Jeremy Howard ở Kaggle, Edd Dumbill, Roger Magoulas và đội ngũ tại O'Reilly Media, và Edward Lazowska. James Cortada đã giúp đỡ nhiều. Cũng xin cảm ơn Ping Li của Accel Partners và Roger Ehrenberg của IA Ventures.

Tại *The Economist*, các đồng nghiệp của tôi đã mang đến những ý tưởng và sự hỗ trợ tuyệt vời. Tôi đặc biệt cảm ơn các biên tập viên của tôi, Tom Standage, Daniel Franklin, và John Micklethwait, cũng như Barbara Beck, người đã biên tập báo cáo đặc biệt "Dữ liệu, Dữ liệu ở Mọi nơi", nó là khởi điểm của cuốn sách này. Henry Tricks và Dominic Zeigler, những đồng nghiệp của tôi ở Tokyo, là những hình mẫu luôn luôn tìm ra điều mới mẻ và diễn đạt nó một cách tuyệt vời. Oliver Morton đã mang đến trí tuệ sắc sảo của mình khi cần thiết nhất.

Hội thảo Toàn cầu Salzburg ở Áo mang đến sự kết hợp hoàn hảo của sự nghỉ ngơi bình dị và sự tìm tòi trí thức đã giúp tôi viết và suy nghĩ. Một hội thảo bàn tròn ở Viện Aspen trong tháng 7 năm 2011 đã mang lại nhiều ý tưởng, mà tôi phải cảm ơn những người tham gia và người tổ chức, Charlie Firestone. Ngoài ra, xin gửi lời cảm ơn của tôi đến Teri Elniski vì sự hỗ trợ to lớn của bà.

Frances Cairncross, Hiệu trưởng Trường Exeter, Oxford, đã cho tôi một nơi yên tĩnh để trú ngụ, cùng sự khích lệ lớn lao.

Sự biết ơn sâu sắc nhất của tôi là giành cho gia đình tôi, những người đồng hành với tôi - hay thường xuyên hơn, với sự vắng mặt của tôi. Cha mẹ, chị em, và những người thân khác của tôi đều xứng đáng được cảm ơn, nhưng tôi dành hầu hết lòng biết ơn của mình cho vợ tôi, Heather, và những đứa con của chúng tôi, Charlotte và Kaz. Không có sự hỗ trợ, khuyến khích và những ý tưởng của họ thì cuốn sách này đã không thể ra đời.

Cả hai chúng tôi xin cảm ơn rất nhiều người đã thảo luận về chủ đề dữ liệu lớn với chúng tôi, rất lâu trước khi thuật ngữ này thậm chí được phổ biến rộng rãi. Đặc biệt, chúng tôi cảm ơn những người tham gia trong những năm qua tại Hội nghị Rueschlikon về Chính sách Thông tin, do Viktor phối hợp tổ chức và nơi Kenn là báo cáo viên.

Chúng tôi đặc biệt cảm ơn Joseph Alhadeff, Bernard Benhamou, John Seely Brown, Herbert Burkert (người giới thiệu chúng tôi với Commodore Maury), Peter Cullen, Ed Felten, Urs Gasser, Joi Ito, Jeff Jonas, Nicklas Lundblad, Douglas Merrill, Rick Murray, Cory Ondrejka, và Paul Schwartz.

VIKTOR MAYER-SCHÖNBERGER

KENNETH CUKIER

Oxford / London, tháng 8 năm 2012

HẾT

Màu sơn nào có thể cho bạn biết một chiếc xe đã qua sử dụng vẫn còn trong tình trạng tốt? Làm thế nào các công chức ở thành phố New York có thể xác định các hố ga nguy hiểm nhất trước khi chúng phát nổ? Và làm thế nào những cuộc tìm kiếm của Google dự đoán được sự lây lan của dịch cúm H1N1?

Chìa khóa để trả lời những câu hỏi này, và nhiều câu hỏi khác, là dữ liệu lớn. "Dữ liệu lớn" để cập đến khả năng đang phát triển của chúng ta để nắm giữ các bộ sưu tập lớn thông tin, phân tích, và rút ra những kết luận đôi khi sâu sắc đáng ngạc nhiên. Lĩnh vực này có thể chuyển vô số hiện tượng – từ giá vé máy bay đến các văn bản của hàng triệu cuốn sách – thành dạng có thể tìm kiếm được, và sử dụng sức mạnh tính toán ngày càng tăng của chúng ta để khám phá những điều chúng ta chưa bao giờ có thể nhìn thấy trước. Trong một cuộc cách mạng ngang tầm với Internet hoặc thậm chí in ấn, dữ liệu lớn sẽ thay đổi cách chúng ta nghĩ về kinh doanh, y tế, chính trị, giáo dục, và sự đổi mới trong những năm tới. Nó cũng đặt ra những mối đe dọa mới, từ sự kết thúc không thể tránh khỏi của sự riêng tư cho đến khả năng bị trừng phạt vì những thứ chúng ta thậm chí còn chưa làm, dựa trên khả năng của dữ liệu lớn có thể dự đoán được hành vi tương lai của chúng ta.

Trong tác phẩm thông tuệ tuyệt vời và gây nhiều ngạc nhiên này, hai chuyên gia hàng đầu giải thích dữ liệu lớn là những gì, nó sẽ làm thay đổi cuộc sống của chúng ta như thế nào, và những gì chúng ta có thể làm để bảo vệ chính mình khỏi các mối nguy hiểm của nó.

DỮ LIỆU LỚN LÀ CUỐN SÁCH LỚN ĐẦU TIÊN VỀ ĐIỀU TO LỚN SẮP DIỄN RA.



Quét QR Code để xem phim minh họa Dữ liệu lớn (phụ đề tiếng Anh/Việt)

"Cuốn sách tuyệt vời này phá vỡ các bí ẩn và sự cường điệu xung quanh dữ liệu lớn. Một cuốn sách phải đọc cho bất cứ ai trong kinh doanh, công nghệ thông tin, chính sách công, tình báo, và y học. Và bất kỳ ai khác chỉ vì tò mò về tương lai."

—John Seely Brown, cựu khoa học gia trưởng của Xerox Corporation, và giám đốc Trung tâm Nghiên cứu Palo Alto của Xerox

"Cuốn sách này chứa đựng những hiểu biết sâu sắc về các cách thức mới để khai thác thông tin và cung cấp một tầm nhìn thuyết phục về tương lai. Nó rất cần thiết cho bất cứ ai sử dụng – hoặc bị ảnh hưởng bởi – dữ liệu lớn."

—Jeff Jonas, thành viên và khoa học gia trưởng IBM, IBM Entity Analytics



facebook.com/
nhaxuatban.tre

ISBN 978-604-1-03186-9
Dữ liệu lớn



Giá: 120.000 đ

nxbtre.com.vn



Tủ sách BOOKBT #303

22/08/2017